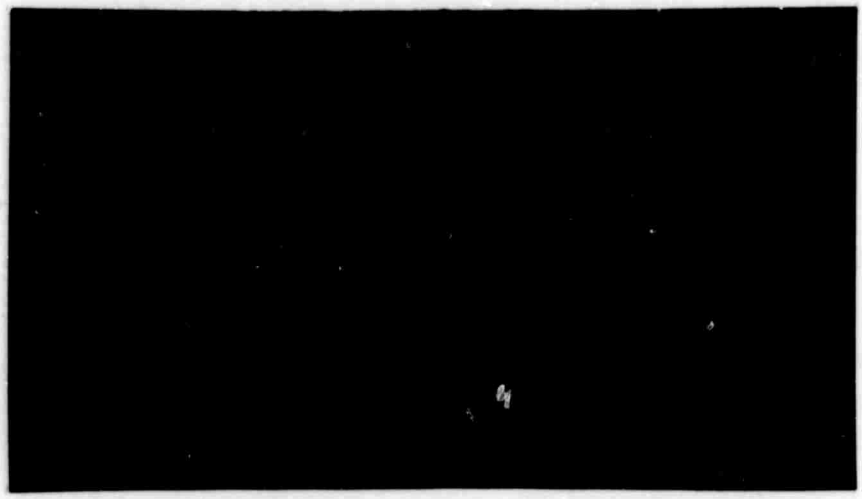
AFOSR 68-2159

AD 672207



1. This document has been approved for public release and sale; its distribution is unlimited.

THE SHUFORD-MASSENGILL CORPORATION

HOW TO SHORTEN A TEST AND INCREASE ITS RELIABILITY AND VALIDITY

EMIR H. SHUFORD, Jr.

THE SECOND SEMIANNUAL TECHNICAL REPORT (WHICH COVERS THE PERIOD NOVEMBER 1966 THROUGH APRIL 1967) OF WORK PERFORMED UNDER CONTRACT NUMBER AF 49 (638) - 1744, ARPA ORDER NUMBER 633, BY THE SHUFORD - MASSENGILL CORPORATION, P.O. BOX 26, LEXINGTON, MASSACHUSETTS, 02173

FOREWORD

This is one in a continuing series of papers concerned with the theory and application of admissible probability measurement techniques and one of a sub-series of papers concerned with the effects of guessing on the interpretation and use of objective test results. This paper constitutes the Second Semiannual Technical Report of work performed in support of the United States Air Force Office of Scientific Research contract number AF 49(638)-1744 sponsored by The Advanced Research Projects Agency of the Department of Defense (ARPA order number 833).

ABSTRACT

Logic and mathematics are employed to yield very conservative estimates of the gains resulting from changing over from choice methods to admissible probability measurement in the administration of existing tests.

Equations and graphs give test reliability and measurement validity as a function of the distribution of ability levels in the population to be tested and as a function of the amount and type of guessing engaged in by this population. Since guessing degrades the performance of choice tests and since the use of admissible probability measurement eliminates guessing, the extent of degradation corresponds to a conservative estimate of the gain resulting from conversion to admissible probability measurement.

In some applications it may be wise to trade off the increase in measurement validity against the advantages of shortening the length of the test. Equations and graphs show how much shorter the new guessing-free test can be and still retain the original measurement validity.

Additional equations and curves show that choice tests with zero measurement validities can have appreciable reliabilities due to differences in guessing strategy in the population.

All the analyses indicate that conversion to admissible probability measurement will yield quite significant improvements in measurement validity along with considerable reductions in test length.

INTRODUCTION

The recent development of the theory of admissible probability measurement (Shuford, Albert & Massengill, 1965) and its successful application in laboratory settings (cf. Toda, 1963; Shuford, 1965) promises to have a profound impact on the theory and practice of objective and semi-objective testing. This new capability to measure an individual's degree of confidence in the correctness of his answers to a test item means that a great deal more information can be obtained about the individual's state of knowledge. The additional information yielded by this new method of testing may be utilized in many different ways. Some of these ways, though promising great benefits and improvements over existing procedures, imply the development of new test items and structures and in other cases creation of new instructional strategies and materials (Shuford, 1965; Shuford & Massengill, 1966). There are however, other more immediately applicable ways of using this new capability to improve the performance of existing testing programs. Since admissible probability measurement can be used with any true-false, multiple-choice or fill-in-the-blank test and since as in the case of conventional choice testing the item scores can be summed to obtain a total test score, it is entirely feasible to change from the conventional choice method over to admissible probability measurement in the administration of an existing test. Such a changeover does not require the writing of any new test items nor does it require the development of new ways of analyzing and utilizing test scores. Can such a simple and minimal change result in any benefits to a testing operation?

Section B of the First Semi-Annual Technical Report (Shuford & Massengill, 1966) considered just this question and arrived at an affirmitive answer. The report took a very conservative approach to estimating the benefits to be derived from substituting admissible probability measurement for choice testing. The approach was conservative in two senses. First, it was assumed that data analysis and personnel decisions would be taken only on the basis of total test score. The use of total test score dominates current personnel measurement and testing practice because of the unreliability of the individual item scores. Adding together these scores from the unreliable items, in effect, builds up the sample size and thus increases the reliability of the total test score. With admissible probability measurement, however, each item score is in a sense completely reliable so that there is no longer the same compulsion to sum item scores to build test reliability. One is tempted rather to look at the pattern of individual item scores to arrive at personnel decisions. But, as stated above this temptation has been temporarily resisted. The potential information inherent in the pattern of item scores will be sacrificed to deal only with total test scores as is the current practice. This is one respect in which the analysis reported in Section B of the First Semi-Annual Technical Report is conservative in estimating the benefits from changing over to admissible probability measurement.

The other respect in which the analysis is conservative is that it is assumed that persons taking the test either know the answer to an item with some

assurance or are relatively uncertain about the answer to an item and that admissible probability measurement will be used *only* to discriminate whether a person knows or is guessing at the answer to an item. The item score obtained with admissible probability measurement typically ranges over the continuum from minus one up to plus one and thus can be used to make fine discriminations in the person's state of knowledge concerning the item in question. This potentially useful information is, however, sacrificed by mapping all the possible scores into two categories: one indicating that the person knows the answer to the item; the other indicating that the person does not know the answer to the item. This is the second respect in which the item analysis reported in Section B of the First Semi-Annual Technical Report is conservative in estimating the benefits of changing from choice testing to admissible probability measurement.

Though these two restrictions assumed by the analysis underestimate the benefits to be obtained they do result in a very important advantage. One can use logic and mathematics in a very straightforward manner for a quantitative study of the effects of guessing on the quality of personnel and counseling decisions. It can be determined just how much guessing degrades the performance of any test by comparing the performance of the guessing-contaminated test with the performance of the test freed of the effect of guessing. The amount of degradation can then be used as an estimate of the gains that will result from eliminating guessing from the test and changing it into a guessing-free test. Since changing over to admissible probability measurement will eliminate the effects of guessing from the test (Massengill & Shuford, 1967) the amount of degradation in the performance of the test due to guessing now becomes the conservative estimate of the gains resulting from substituting admissible probability measurement for choice testing.

Section B of the First Semi-Annual Technical Report used probability theory and decision theory to estimate the effect of guessing on the quality of personnel and counseling decisions in most of the major applications of testing. In attempting to keep the mathematics as realistic as possible, however, a price was paid in that laborious numerical computations were required to obtain any quantitative result. Therefore, the study was restricted to a ten-item test used with a population of just one distribution of ability levels. In this respect the generality of these results is quite restricted. The results are important, however, in that they serve the same purpose as existence proofs in showing something is possible. Specifically they show that guessing at the levels commonly encountered in practice seriously degrades the quality of selection, classification and placement decisions based on total test score. Further, they show that guessing can so seriously degrade the performance of a test used for educational and vocational counseling purposes that it is best to abandon testing for this purpose and just act as though every person had the same average ability level. Additionally and less surprisingly the results show that moderate levels of guessing can seriously degrade the reliability and validity of a test. And finally it is shown that a person's test-wiseness, i.e., whether or not he guesses on a test, largely determines his chances of being successful on the test.

In summary, Section B of the First Semi-Annual Technical Report outlines the methodology for obtaining conservative estimates of the gains from admissible probability measurement and shows that these gains could be of great magnitude in a variety of areas of application. The numerical results were limited however, to a ten-item test used with one population having a certain distribution of ability levels. In some cases it would be useful to extend these results and to increase their generality so that they may be used more effectively to guide decisions on whether or not to change from choice testing to admissible probability measurement. This Second Semi-Annual Technical Report accomplishes just this for the area of test reliability and validity.

RATIONALE

It has been shown both theoretically (Shuford & Massengill, 1966) and empirically (Shuford, 1965) that changing from choice testing to admissible probability measurement can increase the reliability and validity of a test. This increased reliability and validity can be important in two respects. First, the higher the validity of the test the better it is in the sense that it yields better information which in turn implies that better decisions can be made on the basis of the test information. There exist some situations, however, where it is better to accept a shorter but less valid test. The longer the test the greater the "price" that must be paid for the test information. Since the reliability and validity of most tests can be increased by adding on additional items, it is apparent that a decision has been made at least implicitly that the additional validity gained by lengthening the test is not worth the cost of a longer test. Now, when we think of using admissible probability measurement with an existing test we should think also of the possibility of trading off the increased test reliability and validity against the possibility of having a much shorter test with its reliability and validity equal to or greater than that of the original test. The shorter test may, of course, be obtained by using only a sub-set of the items in the original test and thus does not require the writing of new test items but just an elimination of items in the original test. This is an easy change to make in a test and one that should be considered in those cases in which the potential reductions in cost in testing more than offset the potential gains from increasing the validity of the test.

There is another reason for shortening the test over and above just the reduced cost of a shorter test. In some instances, a certain amount of testing time may be available at a constant cost. Here, the leftover testing time may be put to good use by introducing new tests which measure other characteristics of the individual. This new battery of tests may be of much greater "band width" (Cronbach & Gleser, 1965) and may greatly improve the performance of the testing program.

From the above, it should be apparent that there must exist some test which can be made at the same time shorter *and* more valid by the changeover from choice testing to admissible probability measurement. What characterizes

these tests? Are all tests of this sort? How large are the gains that can be expected from changing over from choice testing to admissible probability measurement?

These are the types of questions that will be answered here. As before, logic and mathematics will be employed to yield very conservative estimates of the gains resulting from the use of admissible probability measurement. Equations will be derived which give test reliability and measurement validity as a function of the distribution of the ability levels in the population to be tested and of the amount and type of guessing engaged in by this population. Then additional equations will be derived to show what happens when guessing is eliminated by changing over to admissible probability measurement. These equations will be solved and the results plotted over the complete range of parameter values for the most important cases to be encountered in practice. All of the equations hold for tests of any length.

The formal statement of the testing process is similar to that given in Section B of the First Semi-Annual Technical Report and may be found in Mathematical Appendix A. Briefly, the testing process is based on the independent sampling of test items from a large pool and on the independent sampling of persons from a population which is characterized by a specified distribution of ability levels where ability level is defined as the proportion of items in the pool that the individual knows.

EVERYONE GUESSES

Assume now that each person to be tested knows a certain proportion of the test items in the pool. Let this proportion be represented by p which can, of course, range from zero to one. Different people will know a different proportion of the items and thus differ in ability level with respect to the test under consideration. This distribution of ability levels is represented by a beta distribution defined over the interval from zero to one. The beta distribution has two parameters, a and b, which completely determine its shape and location. The parameters a and b can take on any values greater than zero. The mean of the beta distribution can be obtained from the two parameters, thus $\bar{p} = a/(a+b)$. Also the variance of the beta distribution can be written in terms of these parameters, thus

$$\sigma_p^2 = ab/(a+b)^2(a+b+1).$$

The beta distribution is very flexible and can assume many shapes depending of course upon the particular values of a and b selected.

For all the computations carried out below we will use one of the six different beta distributions shown in Figure 1. Distribution A represents an equal distribution of ability levels over the population to be tested. This rectangular distribution is not likely to be found in practice but is of some interest because it represents an extreme. Distribution B represents the distribution of ability levels for a test of some considerable difficulty; while Distribution C represents a distribution of ability levels for

a test which is rather easy. Distribution D is of considerable interest because it is an approximation of conditions that are often found in practice. There is a symmetric distribution of ability levels with considerable variation over the range. Distributions E and F represent respectively symmetric distributions with less variability in ability level, with Distribution F probably representing the other practical extreme to Distribution A in the sense that in A the ability levels are quite spread out while in F the ability levels are clustered around the average ability level of 1/2.

Though there may be other distributions of ability level of interest these six represent quite a few types and probably place reasonable practical bounds on those that will be encountered in practice. The equations given below can, however, be used to derive the results for any of the infinite number of beta distributions of ability levels, but we will solve them only for these six distributions.

Now, if a person knows the answer to a particular item he will answer it and get it correct. If, on the other hand, he does not know the answer, we assume for the purposes of this section that the person will guess at the answer and that he has a probability, θ, of getting the item correct by chance. The guessing level, θ, can range from zero to 1/2. If $\theta = 0$, then no one is guessing on the test, i.e., if a person doesn't know the answer to an item, he leaves it blank. If $\theta = 1/2$, then the maximum amount of guessing possible is occurring. In a true-false test $\theta = 1/2$ also equals the minimum possible amount of guessing because if there are only two possible answers, the probability of chance success must be 1/2. However, this value of $\theta = 1/2$ also represents a maximum amount of guessing for any other test. For example, in a five alternative multiple-choice test the guessing level may be equal to 1/2 because people may have enough information to exclude three of the five alternatives and just have to guess between the remaining two. Likewise, in a constructed response and fill-in-the-blank test the guessing level may be 1/2 if people tend to think of only two possible answers and have to guess which of the two is right. A guessing level of $\theta = 1/5$ is probably a practical minimum in the sense that for a five alternative multiple-choice test it is the smallest value that θ can assume. That is, if a person has no information with which to discriminate between the five alternatives but must just pick one at random, then his probability of chance success is 1/5. Note that if he does have any information, then his probability of chance success would be much greater than 1/5. Even in constructed-response tests it is unlikely that people are able to think of more than five possible answers and thus achieve a guessing level of less than 1/5. So we will take $\theta = 1/5$ to be a minimum practical guessing level and $\theta = 1/2$ to be a maximum possible guessing level, though we will investigate test reliability and validity over the complete range from $\theta = 0$ (representing what can be achieved by admissible probability measurement) up to $\theta = 1/2$.

ONE ITEM TEST RELIABILITY

Suppose that two test items were selected at random from the pool of test

items and that each of the test items are given to persons selected from a population with distribution of ability levels characterized by the beta distribution with parameters a and b. If a person answers an item correctly, he receives a score of one and if he answers incorrectly, he receives a score of zero. Each of the two items can be considered a separate one-item test and the correlation between these test scores would represent the one-item test reliability. Suppose further, that if a person does not know the answer to an item, he guesses with probability, θ, of success. Now from Mathematical Appendix C we have the equation for a one-item test reliability under the condition that everyone guesses at level θ:

$$(1) \qquad r_{xy}(\theta|a,b) = \frac{(1-\theta)^2\sigma_p^2}{\bar{p}_\theta(1-\bar{p}_\theta)}$$

where $\bar{p}_\theta = (1-\theta)\bar{p} + \theta$. Notice that this one-item test reliability depends upon three parameters, θ, a and b. Examination of this equation reveals unequivocally that the test reliability becomes smaller as the variance, σ_p^2, of the distribution of ability levels becomes smaller. In effect, this variance sets an upper limit on test reliability.

It is not so clear how guessing level θ affects test reliability since θ appears in both the numerator and denominator of the equation. The impact of the guessing level can best be seen by fixing the distribution of ability levels, by specifying a and b, and then solving the equation for different values of θ. The results of such computations are shown in Figure 2 for the six distributions of ability level plotted in Figure 1. The curves in Figure 2 indicate clearly that increasing the guessing level decreases test reliability. Examination of the curves for different distributions of ability level shows that test reliability is related to the size of the variance of the distribution of ability levels. And further, that an asymmetrical distribution (Distribution B and C) interacts with guessing level to effect test reliability. The easy test corresponding to Distribution C is rather slightly affected by guessing level whereas the difficult test represented by Distribution B is greatly affected by increasing the guessing level. All in all, the existence of testing significantly reduces the reliability of a test especially when one considers that .20 is the minimal achievable guessing level in existing choice tests and that .50 is a very commonly encountered guessing level.

N-ITEM TEST RELIABILITY

Assume now that a longer test containing n items is formed by randomly selecting additional items from the pool of test items so that two equivalent n-item tests are obtained and given to persons from the population to be tested. In this case, it is shown in Mathematical Appendix C that the Spearman-Brown prophecy formula type of process can be applied to yield the correlation between these two tests of length n. This is the test reliability for an n-item test and it is dependent upon four parameters, n, θ, a and b.

Thus,

$$(2) \qquad r_{xy}(n,\theta|a,b) = \frac{n r_{xy}(\theta|a,b)}{1 + (n-1) r_{xy}(\theta|a,b)} .$$

Notice that the n-item test reliability is a function of the length of the test, n, and of the one-item test reliability. This means, of course, that by knowing the reliability for a one-item test we can determine the reliability for a test of any length. Equation (2) implies that if the test is made longer and longer (n tends toward infinity), the reliability of the lengthened test will approach one. The test can be made perfectly reliable. This is a traditional result from test theory. If this equation were solved and plotted as a function of n, it would generate a curve which increases by smaller and smaller steps as n increases and asymptotically approaches a reliability of one as shown by two of the curves in Figure 4. This ability to solve for the reliability of tests of any length is very important since it allows us to infer the effect of shortening a test and to compare savings from eliminating guessing.

ONE-ITEM MEASUREMENT VALIDITY

Suppose that just one item is selected from the pool of available items and given to a sample of persons from the population to be tested. Each person will make a score of either zero or one point depending upon whether or not they answered the item correctly. Each person is also characterized by having a certain ability level, p, corresponding to the proportion of items in the complete pool to which he knows the answer. Now, what is the correlation between the test score and this ability level? This correlation would represent the ability of the test to measure the person's ability level and in this sense the measurement validity of the test can be derived as is shown in Mathematical Appendix D. The one-item measurement validity of a test in which everyone is guessing at level θ is

$$(3) \qquad r_{xp}(\theta|a,b) = \frac{(1-\theta)\sigma_p}{\sqrt{\bar{p}_\theta(1-\bar{p}_\theta)}} .$$

As in the case of test reliability, this correlation is a function of θ, a and b, i.e., it is affected both by the guessing level and by the parameters of the distribution of ability levels. In fact, by comparing (3) with (1) it may be seen that (3) is nothing more than the square root of (1). Figure 3 shows how this one-item test measurement validity is affected by different levels of guessing for situations based on the different distributions of ability levels. The effect is quite similar to that obtained for test reliability and it must be so due to the direct relation between the two correlations. Test measurement validity is degraded by the existence of guessing and the degradation will be significant for levels usually encountered in

practice.

N-ITEM TEST MEASUREMENT VALIDITY

Now suppose that we randomly take a sample of n items from the pool of items and give it to people from the population to be tested. Consider the correlation between the total test score and the ability level for this test. As in the case of test reliability, the Spearman-Brown prophecy formula type of process can be used to project the measurement validity of a test of any length and the resulting equation as derived in Mathematical Appendix D is

$$r_{xp}(n,\theta|a,b) = \frac{r_{xp}(\theta|a,b)\sqrt{n}}{\sqrt{1 + (n-1)r_{xy}(\theta|a,b)}} \quad (4)$$

Thus, from knowing the one-item test measurement validity we can obtain the measurement validity for a test of any length. This equation is really just the square root of (2) and thus has quite similar properties. For example, as the sample size increases without bound the measurement validity of the lengthened test approaches one. See Figure 4.

MAXIMUM REDUCTION IN TEST LENGTH POSSIBLE WITHOUT REDUCING EITHER TEST RELIABILITY OR MEASUREMENT VALIDITY.

Since the foregoing equations imply that the existence of guessing reduces both test reliability and measurement validity and that both of these quantities are a function of the length of the test, several interesting results may be deduced. First, and most obviously, if admissible probability measurement were used to eliminate guessing in an existing test then the resulting total test score would be both more reliable and valid. Second, though the elimination of items from this new test would decrease both test reliability and measurement validity, there is generally a range of reduced test lengths over which the guessing-free test will be both more reliable and valid than the longer guessing-contaminated test. There will, of course, also be a range of reduced test lengths over which the guessing-free test will be less reliable and valid than the much longer guessing-contaminated tests. And there will be one unique reduced test length at which the guessing-free test will have essentially the same reliability and validity as the much longer guessing-contaminated test. It is shown in Mathematical Appendix E that this reduced test length, n_0, at which the reliability and validity of the guessing-free test exactly matches that of the original guessing-contaminated test can be obtained by solving

$$n_0 = \frac{n_\theta r_{xy}(\theta|a,b)[1-r_{xy}(0|a,b)]}{r_{xy}(0|a,b)[1-r_{xy}(\theta|a,b)]} \quad (5)$$

where n_θ is the length of the original guessing-contaminated test with one-item test reliability $r_{xy}(\theta|a,b)$ while the one-item test reliability of the

guessing-free test is $r_{xy}(0|a,b)$.

The value $n_\theta - n_0$ is typically a maximum possible reduction in test length in the sense that very seldom would one want to reduce the test reliability and measurement validity below that yielded by the original guessing-contaminated test. Any test length smaller than n_0 would yield a test reliability and measurement validity smaller than that of the original test while a test length greater than n_0 would yield a test reliability and measurement validity larger than that of the original test.

The examination of (5) clearly indicates that maximum reduction in test length possible is a function of the length of the original test n_θ. The dependence of n_0 on the guessing level, θ, occurring in the original test and upon the distribution of ability levels as represented by a and b is not clear from examination of (5). Therefore, we have set $n_\theta = 100$, a fairly typical length for a test used for personnel decisions, and have solved (5) for different values of θ for each of the six distribution of ability levels considered in this report. The results are plotted in Figure 5. Remember that D represents a rather classical distribution of ability levels. In this case, if the guessing level in the original 100-item test were at the minimal value of 1/5 then changing to admissible probability measurement could result in reduction of the length of the test to about a 63-item test, while if guessing were occurring at the maximum level of 1/2, then the admissible probability test could be made as short as 30 items. Somewhat greater savings result in the case of a more difficult test as indicated by Curve B while slightly smaller savings result in the case of easier tests as indicated by Curve C.

The reduction in test length indicated by these curves is not insignificant. The savings are of such magnitude that using admissible probability measurement to eliminate guessing means that now two or three tests can be given to increase a "bandwidth" of the testing program without any increase in total testing time. The increased "bandwidth" could yield a very great improvement in the overall testing process as argued by Cronbach & Gleser (1965).

It should be understood that the use of admissible probability measurement does not require reduction in the length of the test to exactly the amount indicated in these figures. If the reduced test is shorter than that indicated on the curves then it will be less reliable and valid than the original tests, while if it is longer it will be more reliable and valid than the original guessing-contaminated test. The optimal length of the new test should be determined by careful comparisions of the value of increased validity with the value of reduced test length, possibly to increase "bandwidth" of the testing program.

SOME PERSONS GUESS, OTHERS DON'T

TEST RELIABILITY

The situation analyzed above is realistic for many applications of testing

but not for all. To be more explicit, suppose that some of the testing population were test-wise and would guess whenever they did not know the answer to a test item while others in the population to be tested were not test-wise and invariably would choose to skip an item rather than to guess at its answers. This situation is sometimes encountered in testing programs. It is not covered by the analysis given above since there it was assumed that everyone guesses. Here we will assume that a certain proportion, q, of the population to be tested will guess at level θ and that the rest of the testing population, represented by the proportion 1-q, will never guess on the test. This is such a basic change in the description of the testing process that it is quite possible that it will make significant changes in our conclusions. As shown in Mathematical Appendix C the one-item test reliability when a proportion q of the population to be tested guesses at level θ is:

$$r_{xy}(q,\theta|a,b) = \frac{p_{22} - \bar{p}_q{}^2\theta}{\bar{p}_{q\theta}(1-\bar{p}_{q\theta})} \tag{6}$$

where

$$\bar{p}_{q\theta} = q[(1-\theta)\bar{p} + \theta] + (1-q)\bar{p},$$

$$p_{22} = q[(1-\theta)^2\mu^2 + 2\theta(1-\theta)\bar{p} + \theta^2] + (1-q)\mu^2$$

and

$$\mu^2 = \frac{a(a+1)}{(a+b)(a+b+1)} .$$

The n-item test reliability as derived in Mathematical Appendix C is:

$$r_{xy}(n,q,\theta|a,b) = \frac{nr_{xy}(q,\theta|a,b)}{1 + (n-1)r_{xy}(q,\theta|a,b)} . \tag{7}$$

Equation 6 is too complicated to tell by inspection how the different parameters q, θ, a and b affect test reliability. However, by solving the equation for the six distributions of ability levels considered here, we can gain some idea as to how ability level, guessing level and the proportion guessing affects test reliability.

If 1/2 of the population to be tested guesses at level θ while others in the population do not guess then test reliability varies as a function of guessing level as shown in Figure 6. These results are quite different from those obtained and graphed in Figure 2. In Figure 2, increased guessing always lowers test reliability. In contrast to that consistant and neat result, we now find that increased guessing can serve to increase test reliability rather than to reduce it.

Does this strange result of guessing increasing test reliability hold only when 1/2 of the tested population guesses or does it hold for other situations too? Figure 7 examines test reliability for the minimal guessing level of $\theta = 1/5$ for all possible proportions q, all the way from the ex-

treme case of no one in the tested population guesses to the case in which everyone in the tested population guesses. Figure 8 shows the same analysis for the case in which all guessing is done at the maximal level, $\theta = 1/2$. Examination of these two figures shows that the phenomenon is not unique for $q = 1/2$.

These results certainly cast doubt upon the gains to be expected from changing over to admissible probability measurement. To be more specific, if the testing situation is one in which some of the people guess while others don't then changing to admissible probability measurement to eliminate guessing may result in a test of lesser reliability than that of the original test. In such cases, a lengthening of the test may be required to yield the same reliability as that of the original test. Whether a gain or a loss is realized from the changeover to admissible probability measurement depends very critically upon the combination of parameter values approriate to the test situation. Taken together the results of this new analysis suggest that changing over to admissible probability measurement may yield only slight benefits and in many cases will actually impair the reliability of the test.

Before concluding, however, that changing over to admissible probability measurement holds little promise for improving testing it might be worthwhile to take a look at the measurement validity of a test used with a population where some persons guess while others don't.

TEST MEASUREMENT VALIDITY

As shown in Mathematical Appendix D, the one-item measurement validity for such a test is

$$r_{xp}(q,\theta|a,b) = \frac{(1-q\theta)\sigma_p}{\sqrt{\bar{p}_{q\theta}(1-\bar{p}_{q\theta})}} \cdot \tag{8}$$

The n-item measurement validity is

$$r_{xp}(n,q,\theta|a,b) = \frac{r_{xp}(q,\theta|a,b)\sqrt{n}}{1 + (n-1)\,r_{xy}(q,\theta|a,b)} \cdot \tag{9}$$

Figure 9 shows the one-item measurement validity for each of the six distributions of ability level when half the persons guess. Notice that in this case, increasing the guessing level results in decreased test validity. Figure 10 shows one-item measurement validity when a proportion q of the tested population are guessing at a minimal guessing level $\theta = 1/5$. Here notice that increasing the proportion of people guessing decreases test validity. Figure 11 shows the corresponding results when those persons guess-

ing are guessing at a maximum guessing level $\theta = 1/2$. Again increasing the proportion of persons guessing decreases test validity. These results are more in accord with what was found before. The results shown in these figures and the results of other computations indicate that guessing in any amount done by any proportion of the tested population can never increase test measurement validity. This result agrees with intuition much better than the result concerning test reliability.

Look at (8) and notice that the test measurement validity is no longer equal to the square root of the test reliability. This breakdown of the relation between test reliability and validity has a very important implication which can be seen by examining (9). Now, the measurement validity of the test when tests of length greater than one are considered depends both upon the one-item test measurement validity and upon the one-item test reliability which can no longer be expressed in terms of one another. In testing situations where everyone guesses we found that lengthening the test indefinitely made both test reliability and measurement validity approach the maximum possible value of one, that is, resulted in a completely reliable and completely valid test. In this new situation where some guess while others don't, increasing the length of the test without bounds results in the test reliability approaching maximum possible value of one, but the test measurement validity may approach some other value less than one. See Figure 4 for an illustration of this. That is, depending upon the circumstances, it may be impossible to obtain a completely valid test. But the fact that some persons guess while others don't sets an absolute upper limit on the measurement validity that can be yielded by the test no matter how long it is. This upper limit on test measurement validity is

$$r_{xp}(\infty,q,\theta|a,b) = \frac{r_{xp}(q,\theta|a,b)}{\sqrt{r_{xy}(q,\theta|a,b)}} . \tag{10}$$

While the maximum percent of true variance which can be accounted for by the test of infinite length is given by the square of (10). This percent of true variance accounted for is considered a better measure of test performance than test measurement validity.

Figure 12 shows the upper limit in the percent of true variance accounted for even if the test is of infinite length for the case in which 1/2 of the persons guess at level θ for each of our six distributions of ability level. These curves indicate that any amount of guessing degrades the performance of the test. This degradation becomes of some significance for even a minimal guessing level of $\theta = 1/5$ and increases much more by the time the guessing level reaches a maximum of $\theta = 1/2$. Thus, if some persons in the tested population guess while others don't, the use of a conventional choice method which allows for guessing means that there will be a barrier to the maximum performance that can be realized by the test. This barrier can not be breached as long as we continue to use conventional choice methods for test administration.

Figures 13 and 14 show the upper limit of test performance when those persons guessing guess at the minimum level and at the maximum level. These figures make it quite clear that test performance is degraded whenever one encounters a mixed population where some persons guess and others don't. It is much better either to have everyone guessing or no one guessing and these are the only two situations that eliminate the barrier on maximum test performance.

MAXIMUM REDUCTION IN TEST LENGTH POSSIBLE WITHOUT REDUCING MEASUREMENT VALIDITY

So far the results apply only to tests of infinite length. What happens when we consider the more realistic case of using a test of finite length? In particular, we can consider the maximum reduction in test length possible by changing over to admissible probability measurement to eliminate guessing. In this new situation, we will of course get different results depending upon whether we equalize reliabilities or measurement validities. Since measurement validity, not reliability, is the real measure of test performance, we need concern outselves only with measurement validity. If the guessing-free test has the same or greater validity, we don't really care that it has less reliability than the original guessing-contaminated test. This may seem counter-intuitive but the next major section below may provide some understanding of why we shouldn't pay too much attention to test reliability when the test is affected by guessing. As before, we can solve for the reduced test length, n_0, at which the validity of the guessing-free test exactly matches that of the original guessing-contaminated test as shown in Mathematical Appendix E:

$$(11)\quad n_0 = \frac{n_{q\theta} r^2_{xp}(q,\theta|a,b)[1 - r_{xy}(0|a,b)]}{r_{xy}(0|a,b)\{1 - r_{xy}(q,\theta|a,b) + n_{q\theta}[r_{xy}(q,\theta|a,b) - r^2_{xp}(q,\theta|a,b)]\}}$$

where, $r_{xy}(0|a,b)$ is the one-item reliability of the guessing-free version of the test.

This reduced test length, n_0, is shown for a 100-item test given to a population with Distribution D of ability levels some of whom guess at the minimum level in Figure 15 and at the maximum level in Figure 16. Test reliabilities and measurement validities are also shown on these graphs. The measurement validities for the new guessing-free test and for the old guessing contaminated test will of course be the same. The test reliabilities, however, are different with the new guessing free test having somewhat less reliability than the old guessing-contaminated test of greater length.

As to the reduced test length possible, the effect of different guessing strategies is quite dramatic. In the case of minimal guessing, if everyone guesses, the new test can be reduced to about 63 items. If, however, about half of the people guess while the others don't, then the new test can be reduced to about 34 items. In the case of maximal guessing, if everyone

guesses, the new test can be reduced to about 30 items while if about 1/2 the people guess while the others don't, the new test can be reduced to about 5 or 6 items. The existence of differences in guessing strategy in the tested population can greatly degrade test performance and, conversely, can mean that even greater benefits will be yielded by changing over to admissible probability measurement. Reducing the length of a test to 1/20th of its original length or increasing its validity from the high 60's into the high 90's merely by changing the method of test administration is not a trivial benefit.

The same computations have been performed for the other five of the six distributions of ability level. The results are not too different, with greater gains in some cases and somewhat smaller gains in others.

TEST RELIABILITY WHEN MEASUREMENT VALIDITY IS ZERO.

It may be instructive to investigate what happens to test reliability when there is no validity whatsoever to the test. By zero measurement validity we mean that the probability of a person from the tested population knowing an item varies independently from item to item. His ability level is an independent random variable distributed according to the distribution of ability levels. Thus, learning whether a person knows a given item tells you nothing about whether or not he knows the answer to another item. There is no validity whatsoever in such a test.

Now if none of the persons in the tested population were guessing, the test would have no reliability whatsoever. If, however, a proportion q of people consistently guess when they do not know the answer while the rest of the persons refuse to guess, would the rellability still be zero for such a situation? As shown in Mathematical Appendix E the reliability for such a test with zero validity but where a proportion q of persons with a distribution of ability levels of mean $\bar{p}$ are guessing at level θ is

$$r'_{xy}(q,\theta|\bar{p}) = \frac{q(1-q)\theta^2(1-\bar{p})^2}{p_2(1-p_2)} \qquad (12)$$

where $p_2 = q\bar{p}_\theta + (1-q)\bar{p}$.

This equation is not always equal to zero. In fact, it is generally otherwise if there is any guessing whatsoever. Figure 17 shows test reliabilities for a 100-item test as a function of the various levels of guessing. The reliabilities can become quite large even if there is a minimal amount of guessing. Figures 18 and 19 show the same thing for minimal and maximal levels of guessing as the proportion of people adopting a guessing strategy varies. Maximum reliability is observed when about 1/2 the persons guess and 1/2 do not.

Differences in guessing strategies among people can contribute to the reliability of a choice test. In one sense this reliability is completely

spuricus in that the test has no measurement validity. In another sense though, the reliability is not spurious. What is causing this reliability are the differences in guessing strategy of the tested population. This is also what the test is measuring, if it is measuring anything. More specifically the test is measuring the test-wiseness of the tested population. This test-wiseness would also enter into the behavior of the tested population on any other choice test that they might take. Therefore the correlation between any two choice tests would be positive and there would be at least apparent validity between any two choice tests. In some cases, this can be a real validity though it is usually not realized by the user of the test information. These tests are measuring how test-wise people are. This test-wiseness can in turn reflect how much experience persons have had taking tests which in turn can reflect their level of educational attainment, socio-economic background, race and various other factors. So, in this sense, these tests have some validity. However, it would seem to be much more efficient to just have a person fill out a questionnaire directly, stating his level of educational attainment, socio-economic background, race, etc., rather than having this be indicated indirectly through spurious and misleading test results.

It should be realized that if admissible probability measurement were used to administer such a test, guessing would be completely eliminated and zero test reliabilities and measurement validities would be obtained. The test would be shown to be worthless as a test. One wonders how much of the reliability and validity of certain widely used tests is of just this sort. Now, for the first time, we can hope to find out.

SUMMARY AND CONCLUSIONS

Logic and mathematics has been used to deduce the effect of guessing upon test reliability and measurement validity. Analysis shows choice testing to be highly sensitive to the degrading effects of guessing. In fact, the extent of degradation is so unexpectedly large as to be almost unbelievable when viewed in light of the current consensus that the existence of guessing has near trivial effects on the performance of objective and semi-objective tests. It does appear, however, that the logic is inescapable and is in accord with current and meaningful usage in test theory.

Fortunately we are now in a position to resolve this paradox. The logic of decision-theoretic psychometrics promises that we can eliminate the effects of guessing from any test simply by changing over from a choice method to admissible probability measurement in the administration of the test. Then, by carrying out the standard psychometric analyses one can empirically determine the extent to which the original choice test was affected by testing and the benefits resulting from the changeover to admissible probability measurement.

What is the likely result of such empirical studies? Will admissible probab-

ility measurement yield the rather overwhelming benefits indicated by the foregoing analysis? The major hazard to confirmation may be that the mathematics is not complex enough to adequately represent the testing situation. More specifically, the mathematics are based upon the assumption that everybody guesses at the same level. This is in general not true since items may vary in the potency of their misleads and individuals may vary in the extent to which they have partial knowledge of the subject. At considerable cost the mathematics may be generalized to allow for a variable guessing level. It appears that such a generalization would yield rather well-behaved equations which are affected by an *average* guessing level. Thus, the results would fall somewhere between the minimal and maximal guessing levels encountered. While this means that a maximal benefit, say reducing the length of the test to 1/20th of the original size, will not be obtained in practice, neither will a minimal benefit of reducing the length of the test by 1/3 be obtained, but rather something in between, say, reducing the length of the test by a factor of four or five. Such gains still remain of considerable practical importance for many testing applications.

Another respect in which the mathematics may be questioned is that they are based on a single ability level which is constant from item to item. However, whether this is a constant ability level or an average ability level, makes no difference to the analysis as long as one is concerned with total test score. On the other hand, if one is concerned with item analysis, this assumption does make a difference and should be considered.

If it is considered, it can work in favor of changing over to admissible probability measurement. It does it in this way: Some of the analysis given above was concerned with reducing the length of the test to trade off increased reliability and validity against reduced testing time and greater "bandwidth" of the testing battery. These parts of the analysis did not discriminate between test items in the sense that they assumed either that all the items have the same characteristics or that a random sample of items would be selected from the original test. In practice, this restriction does not have to be maintained. To be explicit, one can look at the item characteristics in terms of the data obtained from admissible probability measurement and select the best possible sub-set of items, say the sub-set of items that yield maximum test validity. Such an approach based on an analysis of the structure of the test items will, in general, mean even greater reductions in the length of the test resulting from the conversion to admissible probability measurement. For example, in a test in which the items are completely redundant, such an analysis would indicate that the test can now be reduced down to a one-item test and have the same validity as the original test of any length. This is admittedly an extreme case, but it shows what is possible by analyzing the structure of test items and selectively choosing the items to be included in the revised guessing-free test.

So, all in all, it appears that the existence of guessing has seriously degraded the performance of objective and semi-objective test programs. A new method of test administration, admissible probability measurement,

promises to eliminate the effect of guessing. This holds out the hope that testing programs may be greatly improved and that new and exciting uses may be found which will greatly enlarge the scope of application of objective and semi-objective tests.

REFERENCES

Cronbach, L. J. & Gleser, G. C. (1965) *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Massengill, H. E. & Shuford, E. H., Jr. (1967) What pupils and teachers should know about guessing. Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, E. H., Jr. (1965) Cybernetic testing. ESD-TR-67-229, Electronics Systems Division, L. G. Hanscom Field, Bedford, Massachusetts.

Shuford, E. H., Jr., Albert, A. & Massengill, H. E. (1966) Admissible probability measurement procedures. *Psychometrika*, 31:125-145.

Shuford, E. H., Jr. & Massengill H. E. (1966) Decision-theoretic psychometrics: an interim report. (First Semi-Annual Technical Report) Lexington, Massachusetts: The Shuford-Massengill Corporation.

Toda, M. (1963) Measurement of subjective probability distribution. ESD-TDR•63-407, Electronic Systems Division, L. G. Hanscom Field, Bedford, Massachusetts.

BLANK PAGE

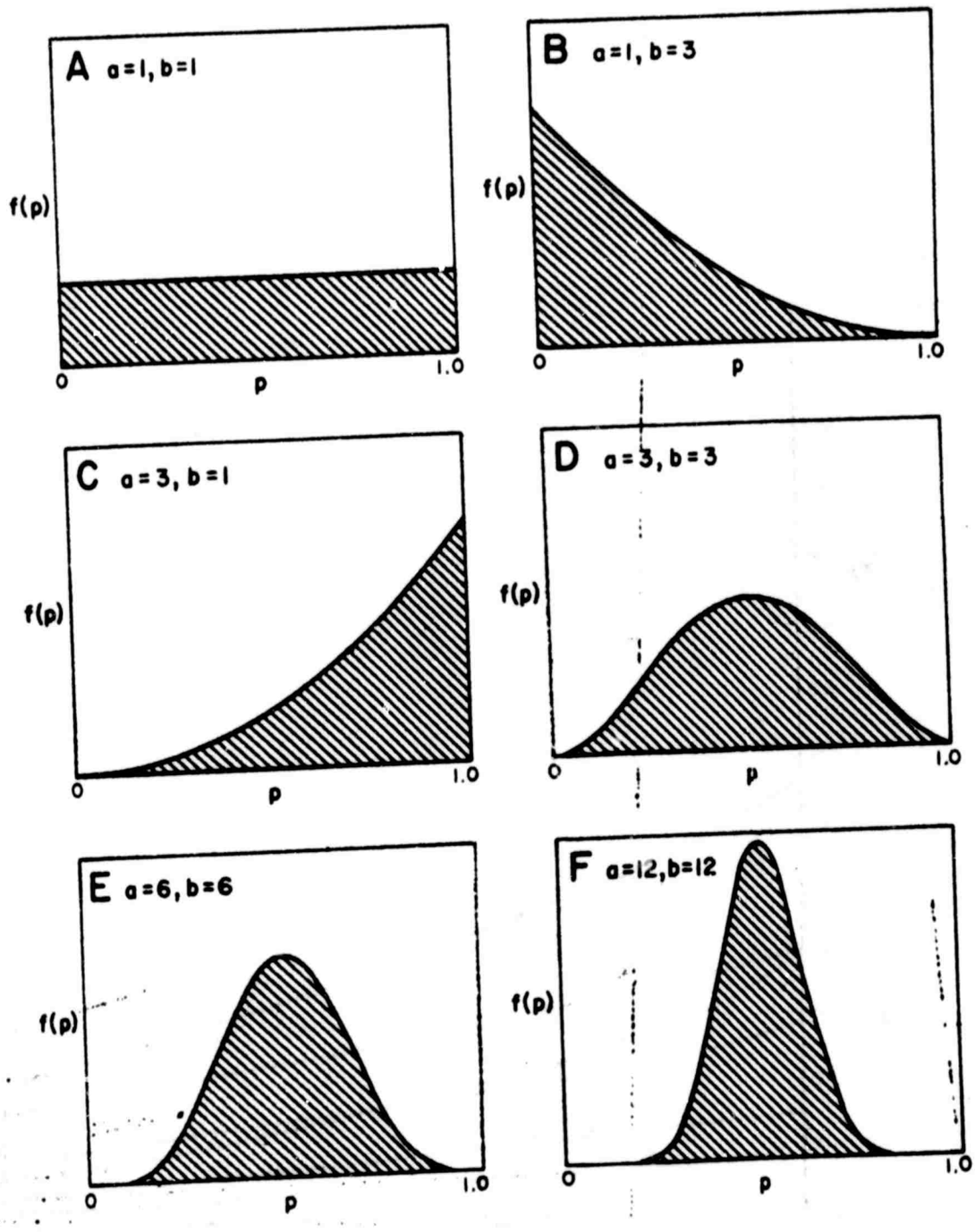


Figure 1

SIX TYPICAL BETA DISTRIBUTIONS OF ABILITY LEVEL

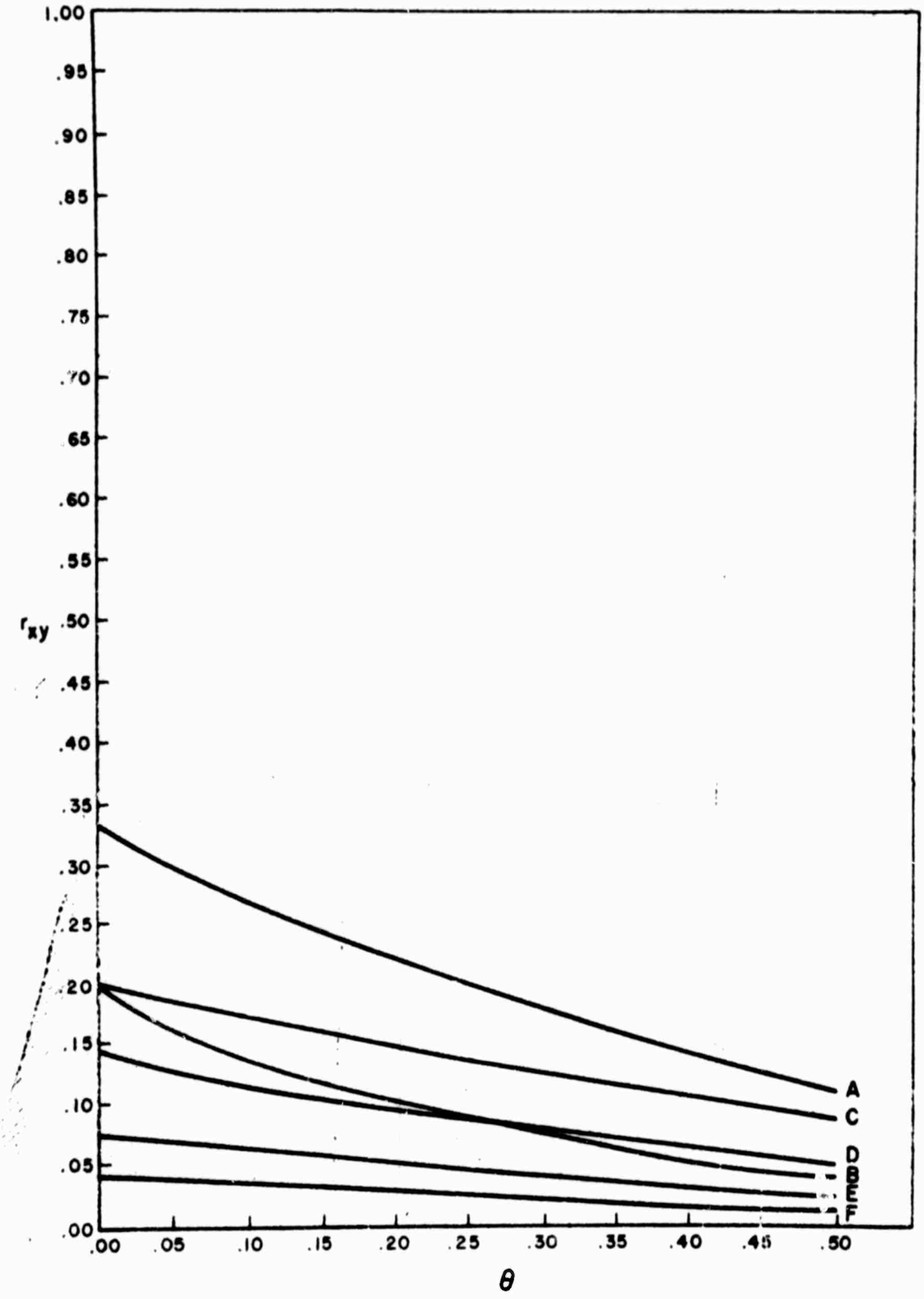


Figure 2

TEST RELIABILITY (ONE-ITEM) WHEN EVERYBODY GUESSES AT LEVEL θ

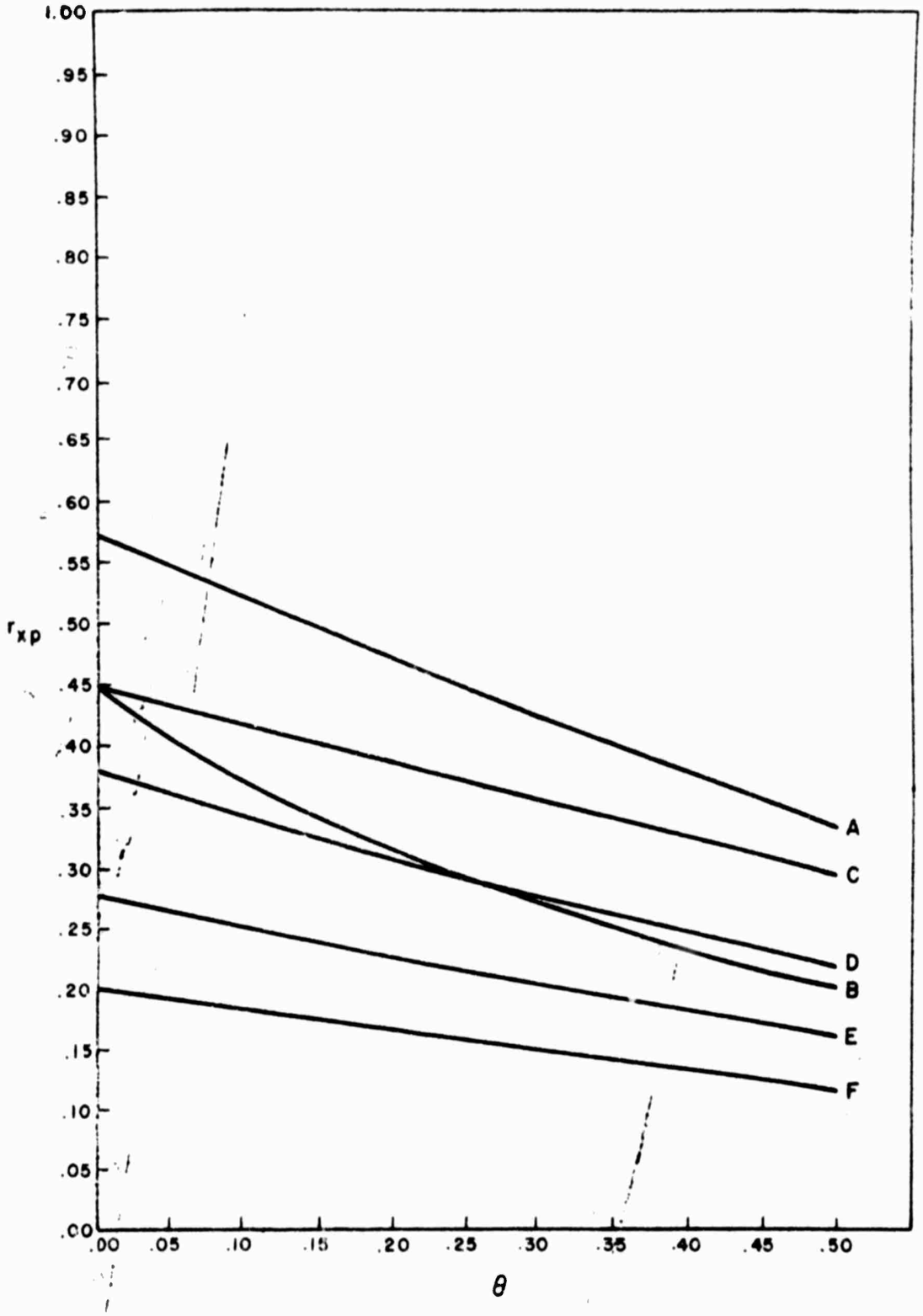


Figure 3

TEST MEASUREMENT VALIDITY (ONE-ITEM) WHEN EVERYBODY GUESSES AT LEVEL θ.

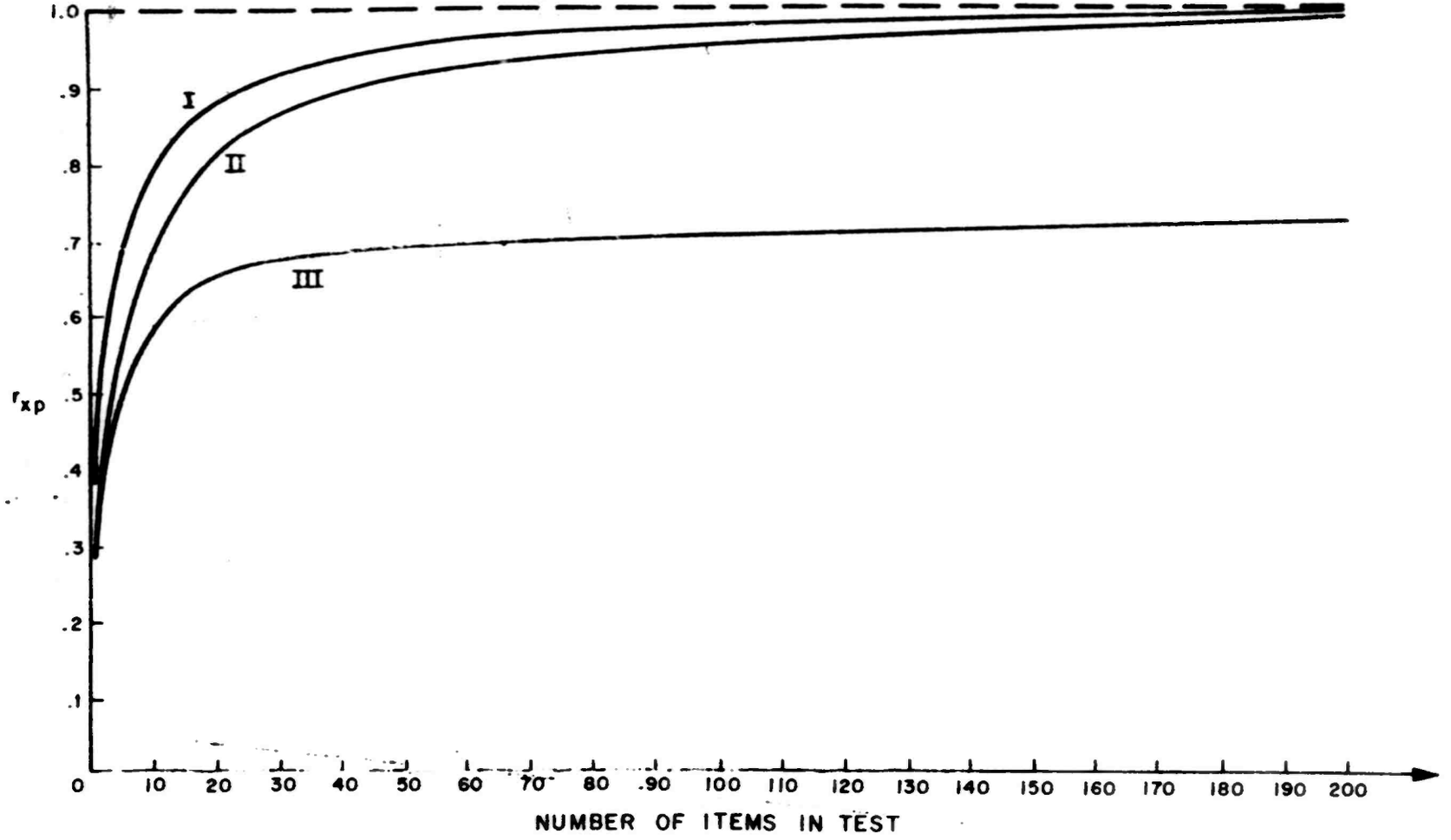


Figure 4

EXAMPLES OF THE GROWTH OF MEASUREMENT VALIDITY
AS TEST IS LENGTHENED

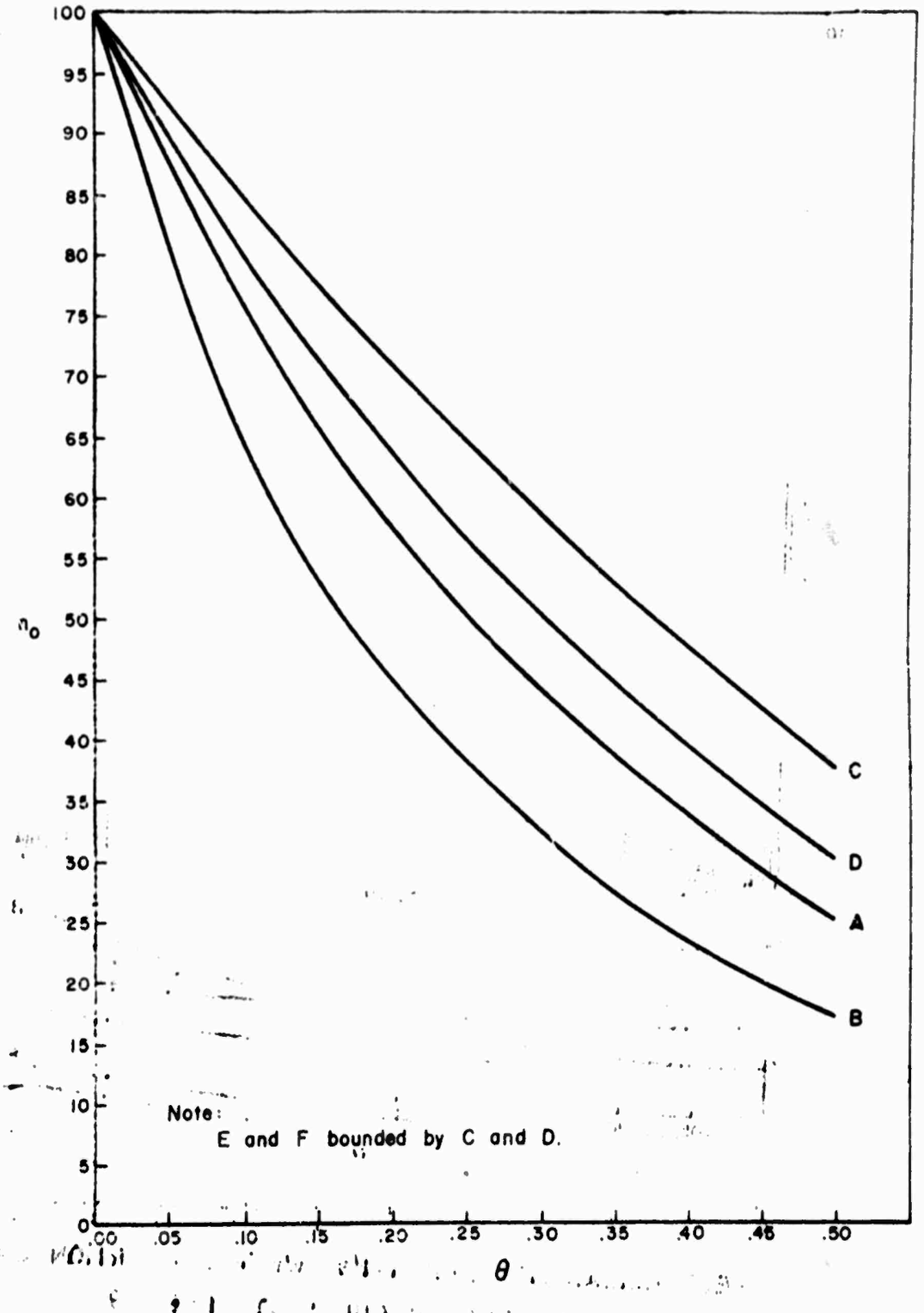


Figure 5

REDUCED TEST LENGTH POSSIBLE WHEN EVERYBODY GUESSES AT LEVEL θ

BLANK PAGE

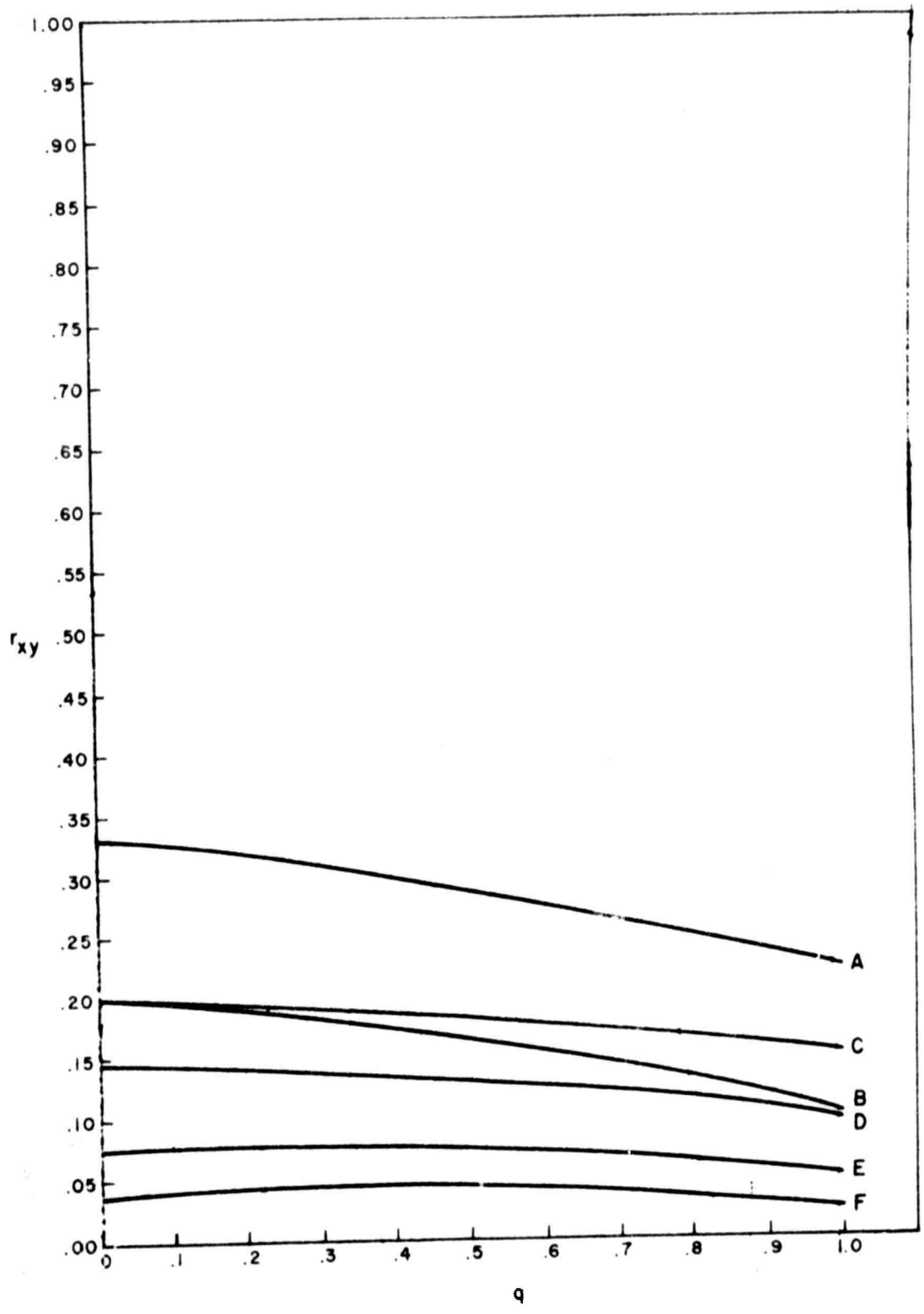


Figure 7

TEST RELIABILITY (ONE-ITEM)
WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{6}$

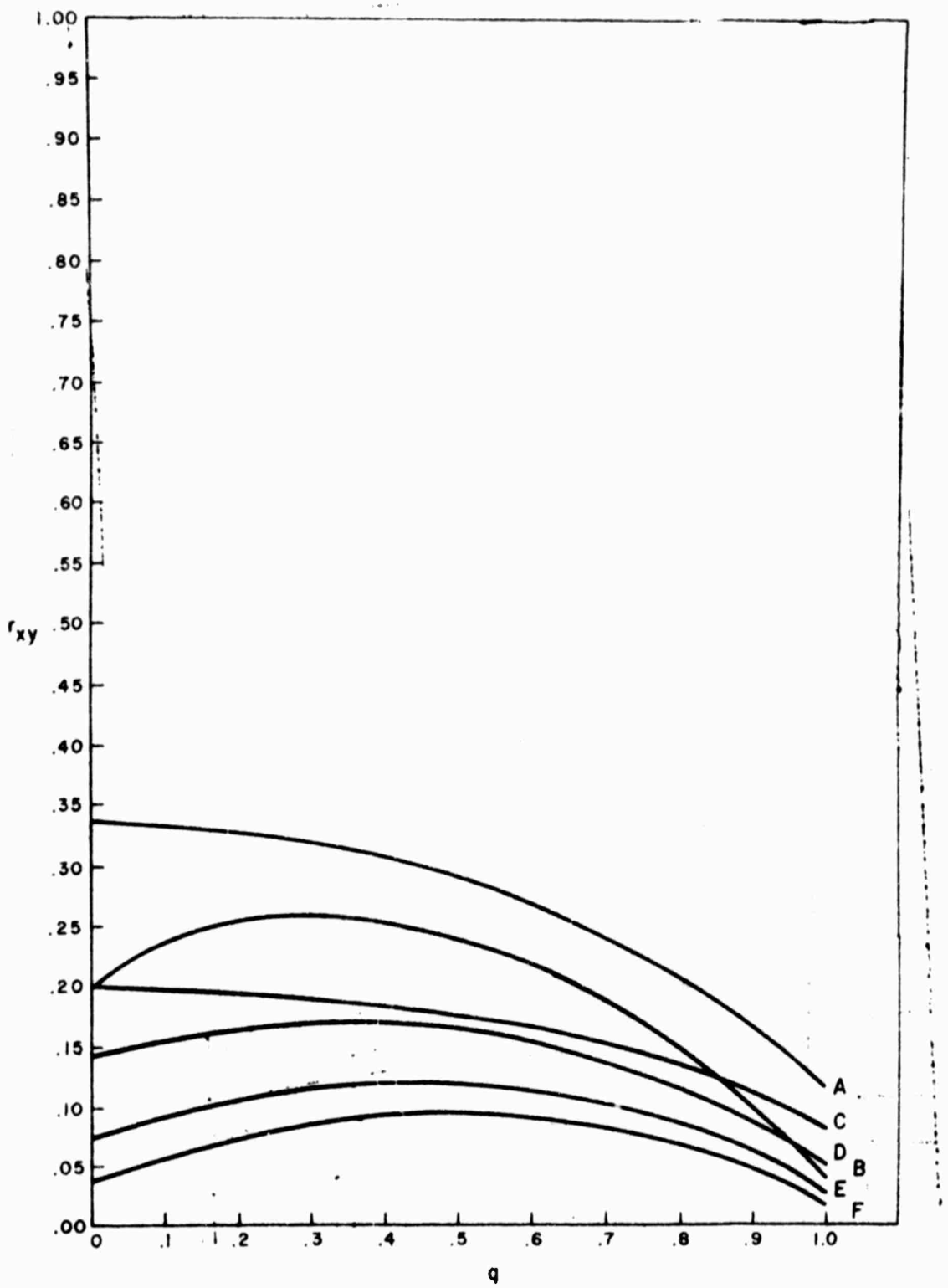


Figure 8

TEST RELIABILITY (ONE-ITEM)
WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{2}$

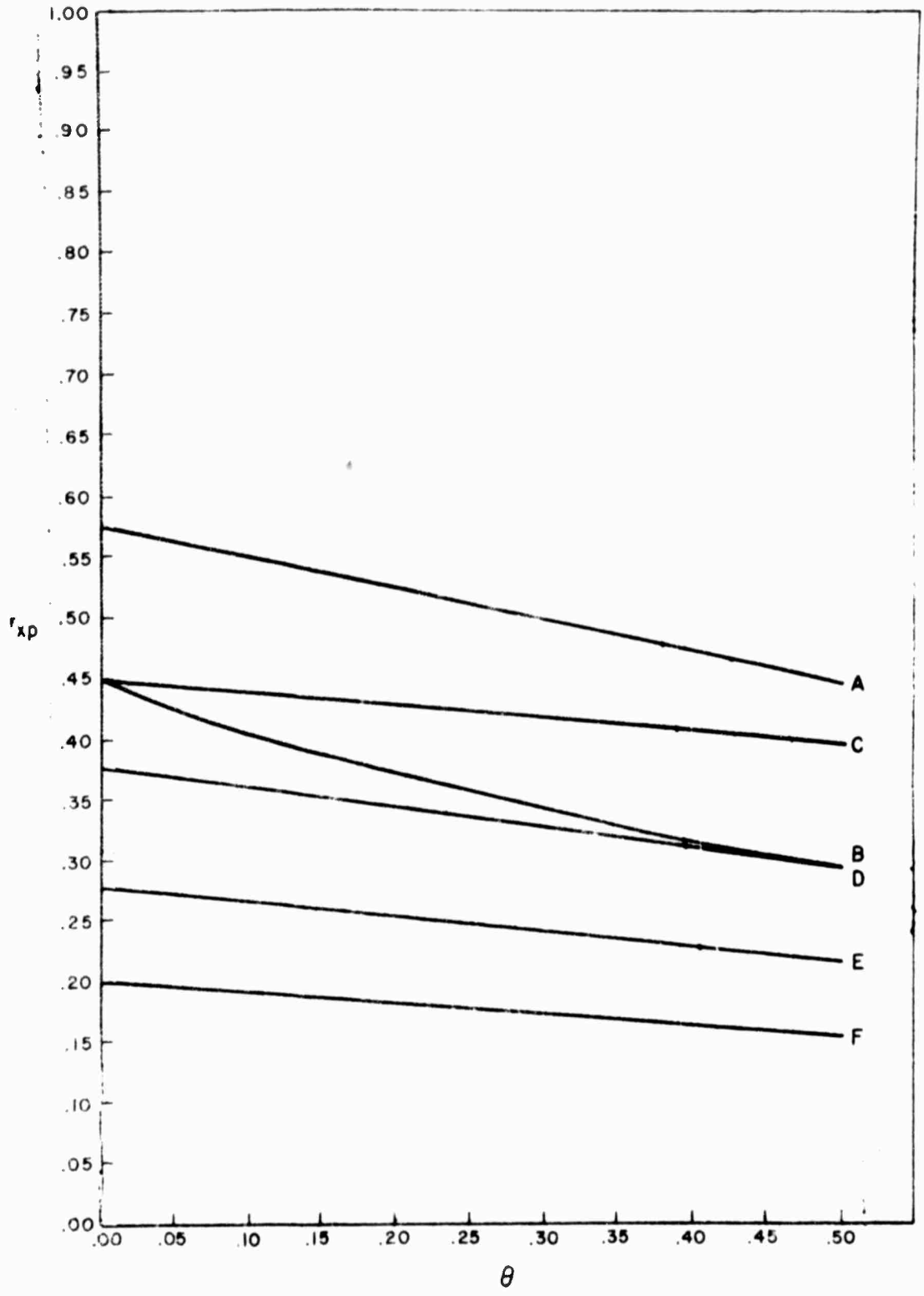


Figure 9

TEST MEASUREMENT VALIDITY (ONE-ITEM) WHEN PROPORTION $q = \frac{1}{2}$ OF PERSONS GUESS AT LEVEL θ

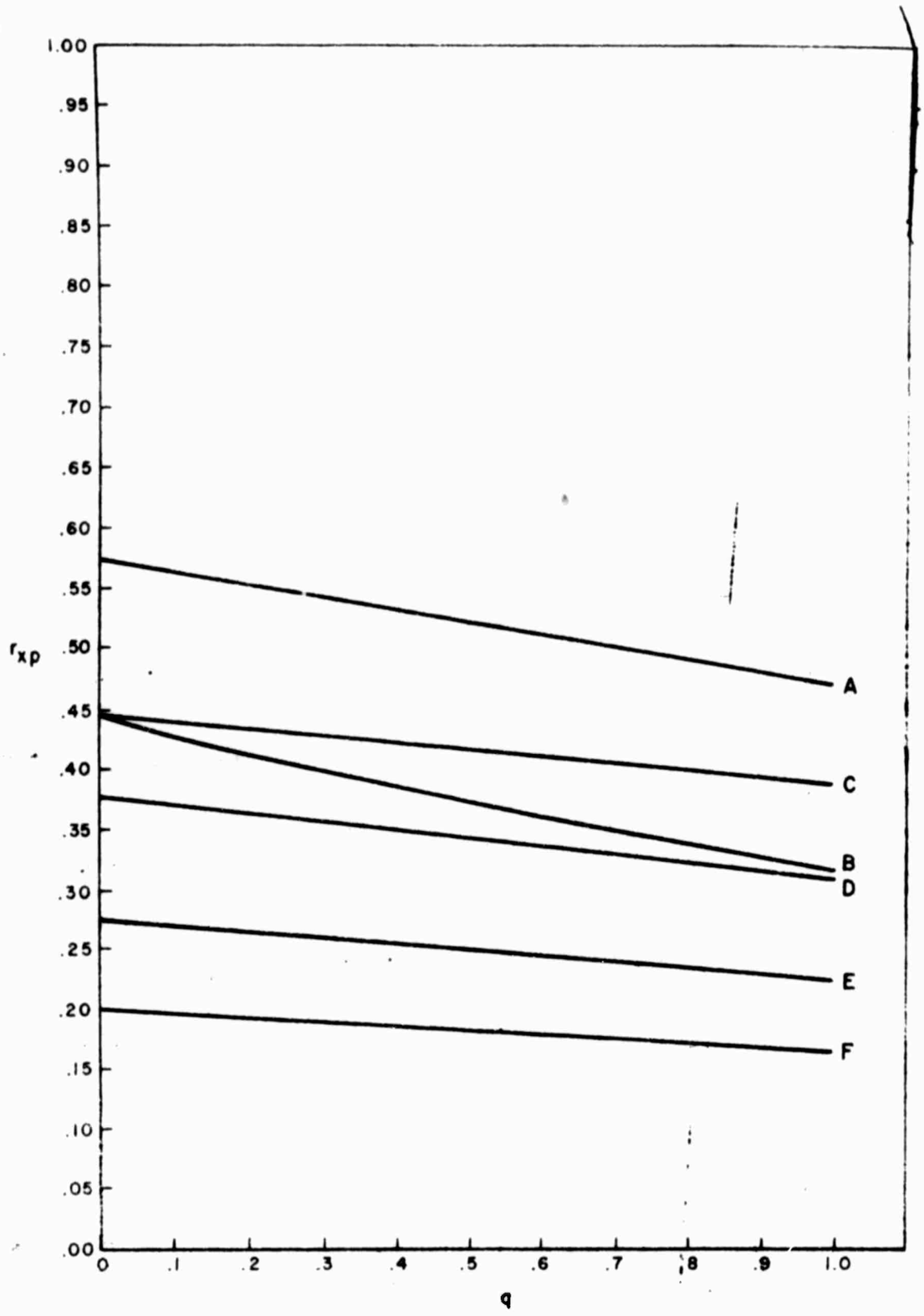


Figure 10

TEST MEASUREMENT VALIDITY (ONE-ITEM) WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{5}$

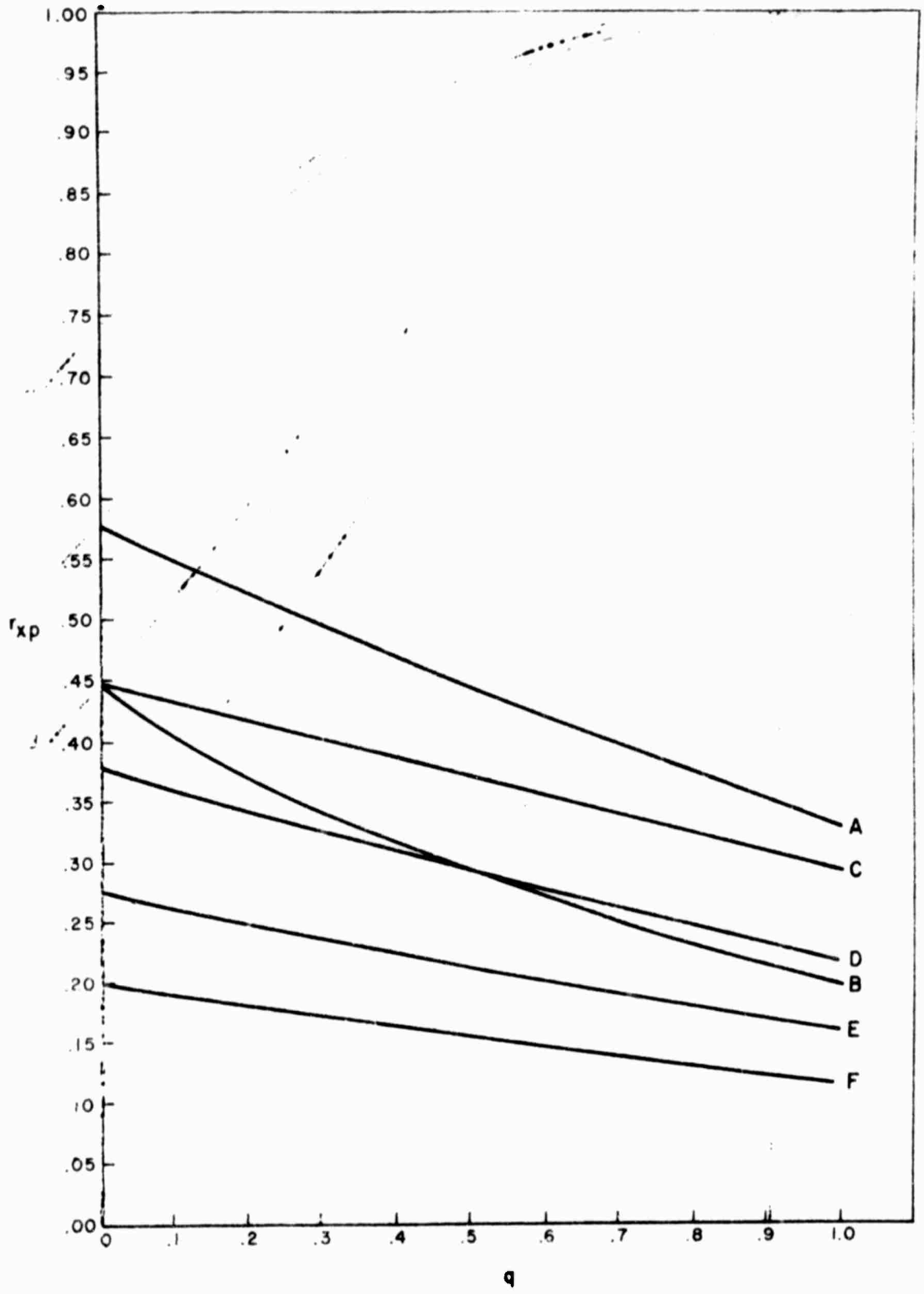


Figure 11

TEST MEASUREMENT VALIDITY (ONE-ITEM) WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{2}$

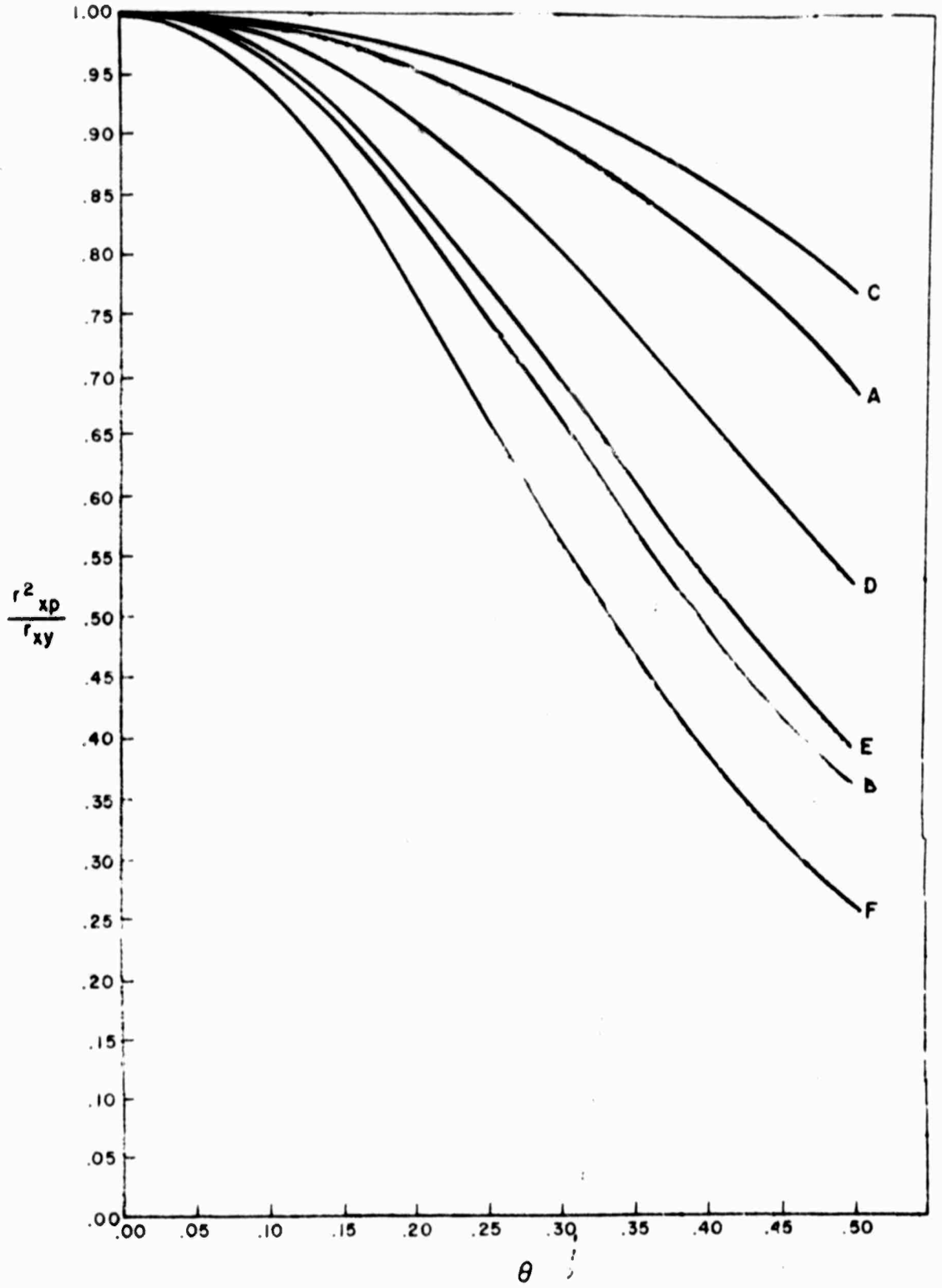


Figure 12

MAXIMUM PERCENT OF TRUE VARIANCE ACCOUNTED FOR (TEST OF INFINITE LENGTH) WHEN PROPORTION $q = \frac{1}{2}$ OF PERSONS GUESS AT LEVEL θ

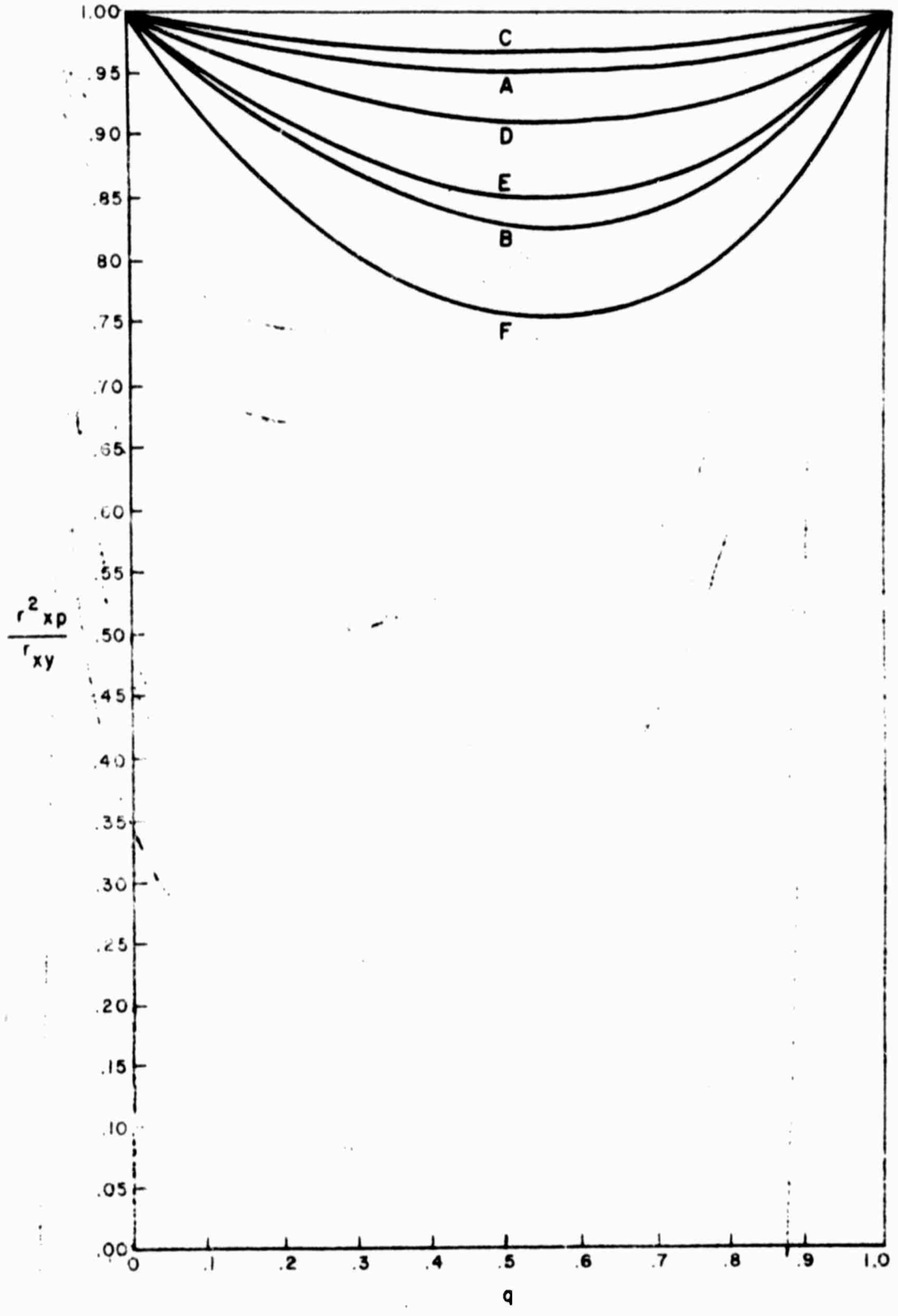


Figure 13

MAXIMUM PERCENT OF TRUE VARIANCE
ACCOUNTED FOR (TEST OF INFINITE LENGTH)
WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{5}$

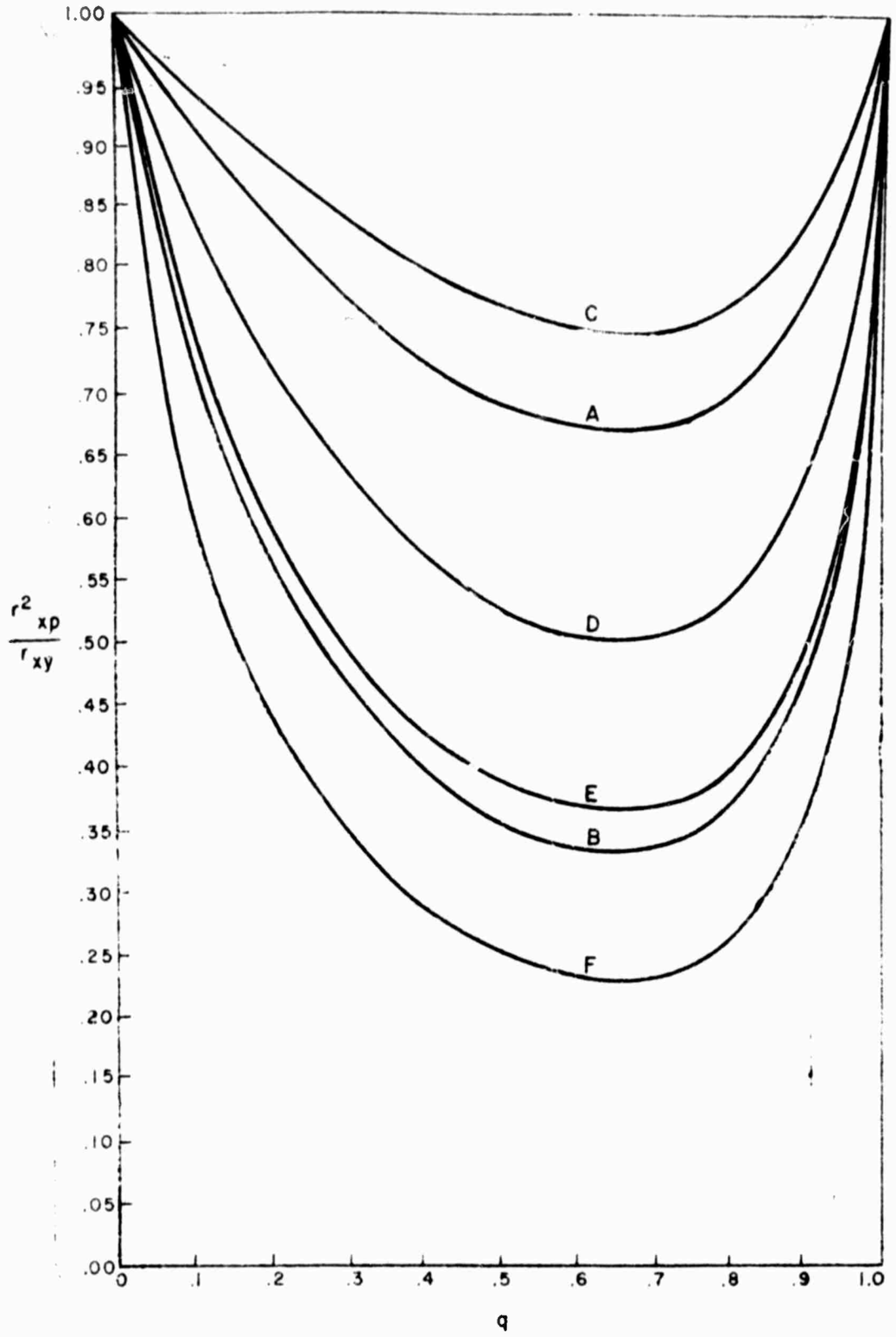


Figure 14

MAXIMUM PERCENT OF TRUE VARIANCE ACCOUNTED FOR (TEST OF INFINITE LENGTH) WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{2}$

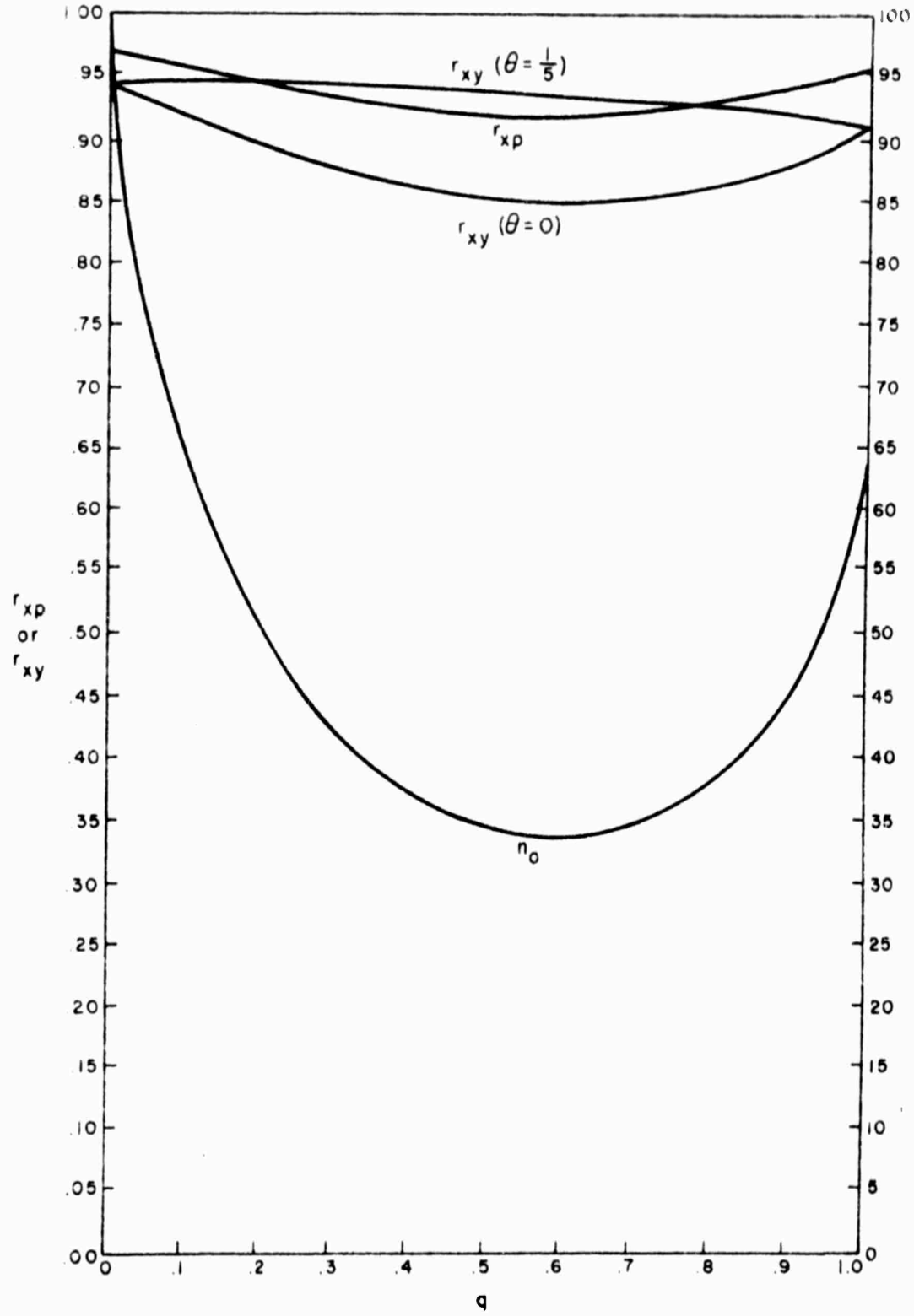


Figure 15

REDUCED TEST LENGTH POSSIBLE FOR DISTRIBUTION D WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{5}$

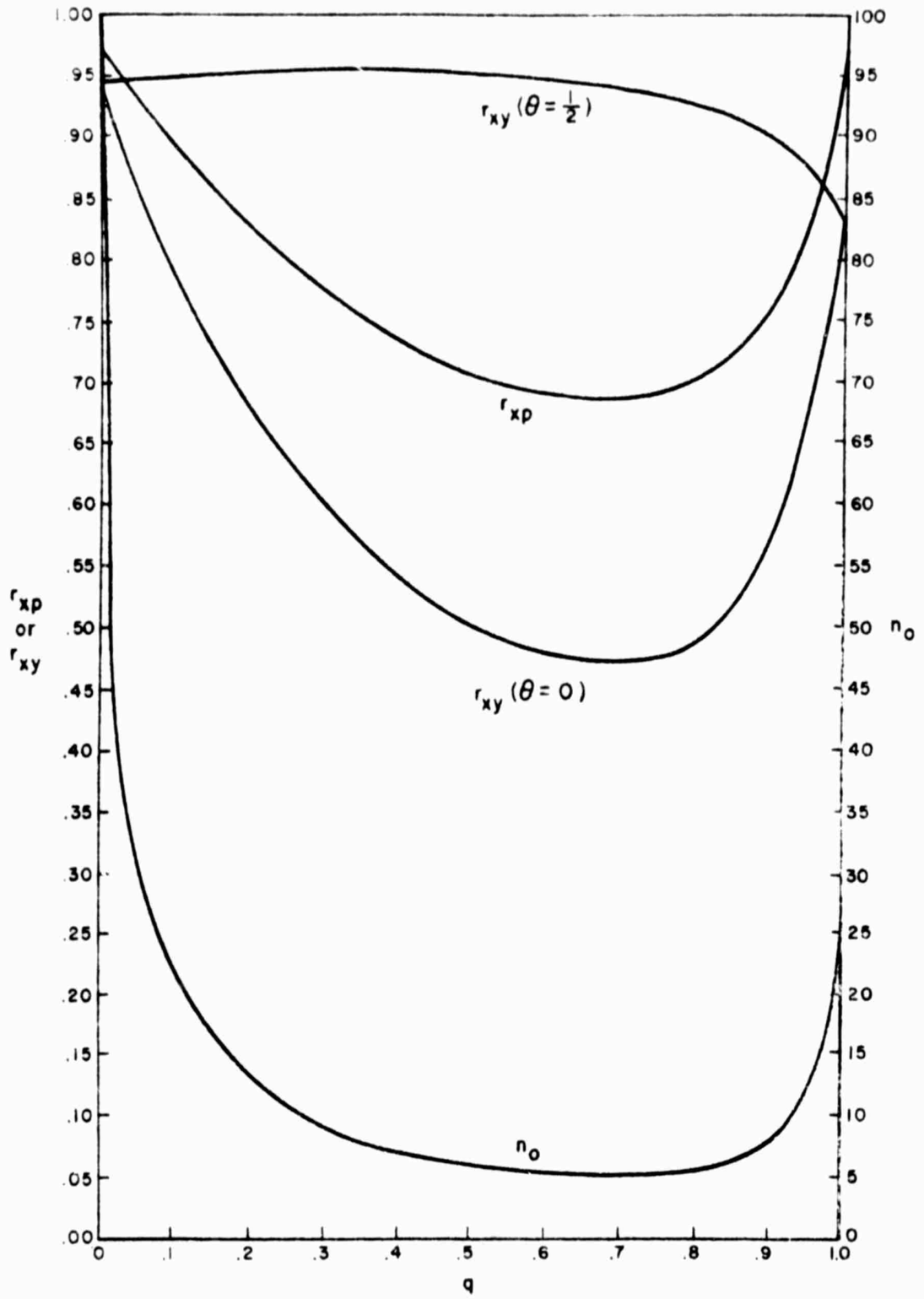


Figure 16

REDUCED TEST LENGTH POSSIBLE FOR DISTRIBUTION D WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta=\frac{1}{2}$

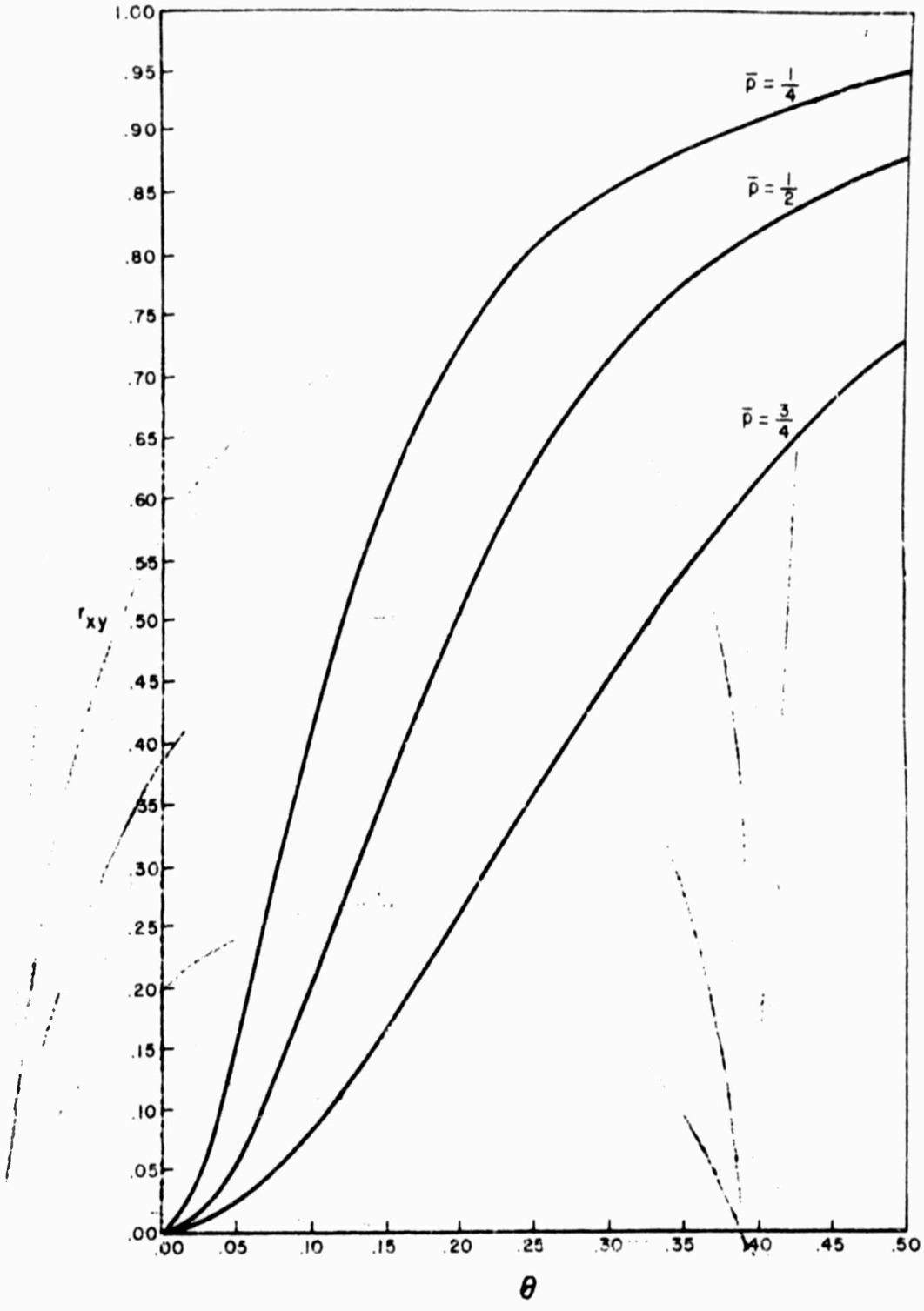


Figure 17

100-ITEM TEST RELIABILITY WHEN PROPORTION $q = \frac{1}{2}$ OF PERSONS GUESS AT LEVEL θ FOR A TEST WITH ZERO MEASUREMENT VALIDITY

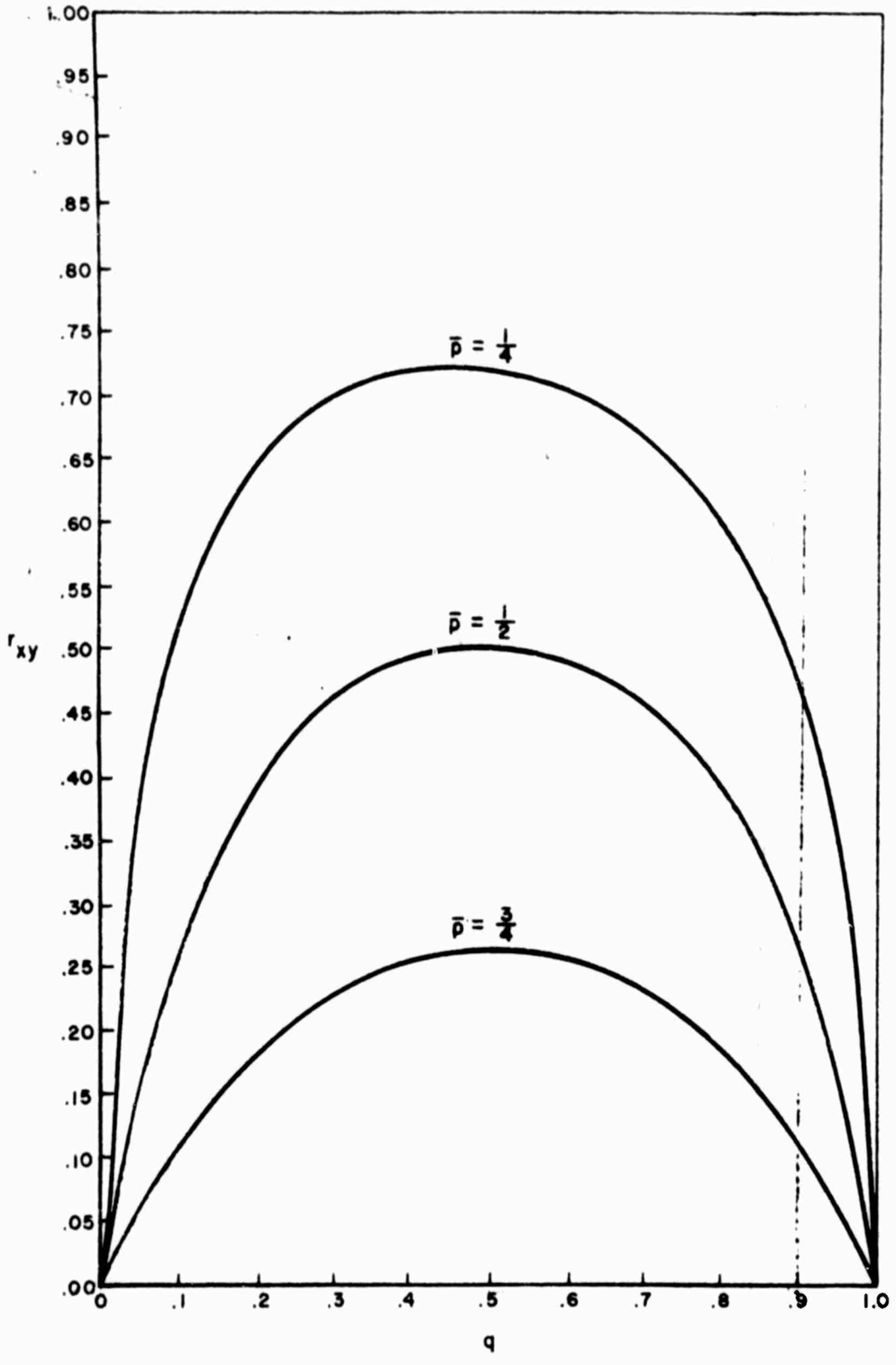


Figure 18

100-ITEM TEST RELIABILITY WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{5}$ FOR A TEST WITH ZERO MEASUREMENT VALIDITY

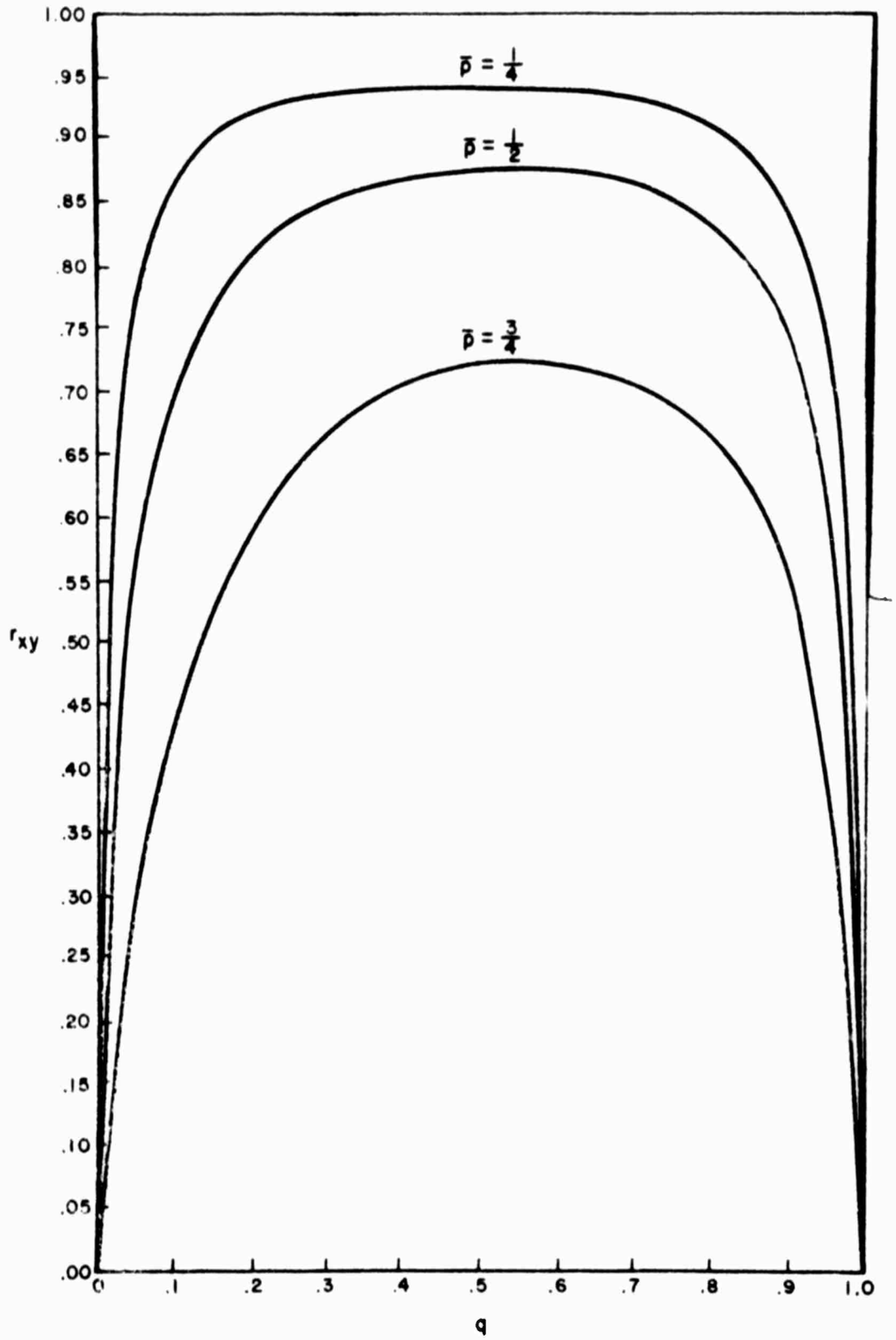


Figure 19

100-ITEM TEST RELIABILITY WHEN PROPORTION q OF PERSONS GUESS AT LEVEL $\theta = \frac{1}{2}$ FOR A TEST WITH ZERO MEASUREMENT VALIDITY

BLANK PAGE

MATHEMATICAL APPENDICES

A. FORMAL STATEMENT OF THE TESTING PROCESS

Define:

$q \equiv$ proportion of persons who guess

$\theta \equiv$ probability of success if person does guess

$p \equiv$ proportion of items in pool that person knows

$X \equiv$ a random variable yielding "1" if person answers item correctly, "0" otherwise.

$Y \equiv$ a random variable yielding "1" if same person answers another item correctly, "0" otherwise.

$Z \equiv$ a random variable, independent of p, yielding "1" if person guesses, "0" otherwise.

Test items are randomly and independently selected from large pool of items.

Persons are randomly selected from relevant population for administration of tests.

B. BETA DISTRIBUTIONS OF ABILITY LEVEL

Since p is the proportion of items in the test pool that a given individual knows, it can serve as a comprehensive measure of his level of ability or achievement. It is, at least, superior to the notion of true score or ability level of classical test theory based upon the number of items answered correctly. This, of course, would be contaminated by the effects of guessing.

Since p has a finite range from zero to one, it is not natural to use a normal distribution to represent the frequency of occurrence of p in a population. The most commonly used and flexible distribution for a variable defined over the interval [0,1] is the beta distribution:

$$f_{\beta}(p|a,b) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$$

where $a,b > 0$.

C. TEST RELIABILITY

One-item test reliability. We need to know the joint and marginal probabilities for the random variables X and Y. Thus,

		X = "0"	X = "1"	
Y	"0"	p_{11}	p_{12}	$p_{1\cdot}$
	"1"	p_{21}	p_{22}	$p_{2\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 2}$	1

Decomposing the joint probabilities:

$$p_{ij} = \Pr(X=i, Y=j, Z=i) + \Pr(X=i, Y=j, Z=0)$$

$$= \Pr(X=i, Y=j|Z=1)q + \Pr(X=i, Y=j|Z=0)(1-q).$$

By further decomposition and by integrating $f_\beta(p|a,b)$ over p we obtain

(C.1) $$p_{22} = q[(1-\theta)^2\mu^2 + 2\theta(1-\theta)\bar{p} + \theta^2] + (1-q)\mu^2$$

(C.2) $$p_2 = q[(1-\theta)\bar{p} + \theta] + (1-q)\bar{p}.$$

Since

(C.3) $$r_{xy} = \frac{p_{22} - p_2^2}{p_2 - p_2^2},$$

we obtain (6) by substituting (C.1) and (C.2) into (C.3) and renaming p_2 as $\bar{p}_{q\theta}$ and we obtain (1) by setting q = 1 and simplifying.

n-item test reliability. Let $X^{(n)}$ and $Y^{(n)}$ represent two n-item tests obtained by sampling from the large pool of items. Then

$$\sigma^2_{X(n)+Y(n)} = \sigma^2_{X(n)} + \sigma^2_{Y(n)} + 2r_{X(n)Y(n)}\sigma_{X(n)}\sigma_{Y(n)}$$

and

(C.4) $$r_{X(n)Y(n)} = \frac{\sigma^2_{X(n)+Y(n)} - \sigma^2_{X(n)} - \sigma^2_{Y(n)}}{2\sigma_{X(n)}\sigma_{Y(n)}}.$$

Upon expanding the right hand side of (C.4) down to the level of individual items realizing that all item variances are equal due to the nature of the basic testing process and that all item inter-correlations are equal to r_{xy}, (C.4) through simplification becomes either (2) or (7) depending upon the definition of r_{xy}.

D. MEASUREMENT VALIDITY

One-item measurement validity. By the definition of correlation,

$$r_{xp} = \frac{\text{Cov } Xp}{\sqrt{\text{Var } X}\ \sqrt{\text{Var } p}} \tag{D.1}$$

Now,

$$\text{Cov } Xp = E(Xp) - E(X)E(p) \tag{D.2}$$

After integration over p, we obtain:

$$E(Xp) = q[(1-\theta)\mu^2 + \theta\bar{p}] + (1-q)\mu^2 \tag{D.3}$$

$$E(X) = q[(1-\theta)\bar{p} + \theta] + (1-q)\bar{p} \tag{D.4}$$

$$E(p) = \bar{p}. \tag{D.5}$$

Substituting (D.3), (D.4), and (D.5) in (D.2) and simplifying we obtain:

$$\text{Cov}(Xp) = (1-q\theta)\sigma_p^2. \tag{D.6}$$

By definition,

$$\text{Var}(X) = E(X^2) - [E(X)]^2 \tag{D.7}$$

and

$$\text{Var}(p) = E(p^2) - [E(p)]^2 . \tag{D.8}$$

Now,

$$E(X) = E(X^2) = q[(1-\theta)\bar{p} + \theta] + (1-q)\bar{p}, \tag{D.9}$$

$$E(p) = \bar{p}, \tag{D.10}$$

and

$$E(p^2) = \mu^2. \tag{D.11}$$

Substituting (D.9) into (D.7) and (D.10) and (D.11) into (D.8) and the results along with (D.2) into (D.1) and simplifying, we obtain

$$r_{xp}(q,\theta|a,b) = \frac{(1-q\theta)\sigma_p}{\sigma_p\sqrt{\bar{p}_{q\theta}(1-\bar{p}_{q\theta})}}$$

which simplifies to (8) and upon setting q = 1, to (3).

n-item measurement validity. As before, let $X^{(n)}$ represent an n-item test obtained by sampling from the large pool of items and let p, of course, represent ability level. Then,

$$\sigma^2_{X(n)+p} = \sigma^2_{X(n)} + \sigma^2_p + 2r_{X(n)p}\,\sigma_{X(n)}\sigma_p$$

and

$$r_{X(n)p} = \frac{\sigma^2_{X(n)+p} - \sigma^2_{X(n)} - \sigma^2_p}{2\,\sigma_{X(n)}\sigma_p} \tag{D.12}$$

Upon expanding the right hand side of (D.12) down to the level of individual items and realizing that all item variances, σ^2_x; item intercorrelations, r_{xy}; and item validities, r_{xp}, are equal, (D.12) becomes

$$r_{X(n)p} = \frac{2nr_{xp}\sigma_x\sigma_p}{2\sqrt{n}\,\sigma_x\sigma_p\sqrt{1 + (n-1)r_{xx}}}$$

which simplifies to

$$r_{X(n)p} = \frac{r_{xp}\sqrt{n}}{\sqrt{1 + (n-1)r_{xx}}} \tag{D.13}$$

This final equation (D.13) then becomes either (4) or (9) depending upon the identification of r_{xp} and r_{xx}.

E. TEST RELIABILITY WHEN VALIDITY IS ZERO

Let p be independent from item to item and proceed as in Appendix C to obtain

$$p'_{22} = q[(1-\theta)\bar{p}_x + \theta][(1-\theta)\bar{p}_y + \theta] + (1-q)\bar{p}_x\bar{p}_y$$

$$= q\bar{p}_\theta^2 + (1-q)\bar{p}^2$$

and

$$p'_2 = q\bar{p}_\theta + (1-q)\bar{p}$$

which, when substituted into (C.3) yields (12).

The procedures of Appendix D may be followed to obtain the validity of such a test. Notice that

$$E(Xp) = \bar{p}\{q[(1-\theta)\bar{p} + \theta] + (1-q)\bar{p}\}$$

$$= E(p)E(X)$$

so that

$$\text{Cov } Xp = E(Xp) - E(X)E(p)$$

$$= E(X)E(p) - E(X)E(p)$$

$$= 0.$$

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

The Shuford-Massengill Corporation Box 26, Lexington, Mass. 02173	UNCLASSIFIED GROUP

HOW TO SHORTEN A TEST AND INCREASE ITS RELIABILITY AND VALIDITY

Scientific INTERIM

Emir H. Shuford, Jr.

REPORT DATE	TOTAL NO. OF PAGES	NO. OF REFS
May 1967	41	6

CONTRACT OR GRANT NO	ORIGINATOR'S REPORT NUMBER(S)
AF 49(638)-1744 PROJECT NO 9720F - 9719 61545C1R 681313	SMC R-9

OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

AFOSR 68-2159

DISTRIBUTION STATEMENT

1. This document has been approved for public release and sale; its distribution is unlimited.

SUPPLEMENTARY NOTES	SPONSORING MILITARY ACTIVITY
TECH, OTHER	Air Force Office of Scientific Research 1400 Wilson Boulevard (SRLB) Arlington, Virginia 22209

Logic and mathematics are employed to yield very conservative estimates of the gains resulting from changing over from choice methods to admissible probability measurement in the administration of existing tests.

Equations and graphs give test reliability and measurement validity as a function of the distribution of ability levels in the population to be tested and as a function of the amount and type of guessing engaged in by this population. Since guessing degrades the performance of choice tests and since the use of admissible probability measurement eliminates guessing, the extent of degradation corresponds to a conservative estimate of the gains resulting from conversion to admissible probability measurement.

In some applications it may be wise to trade off the increase in measurement validity against the advantages of shortening the length of the test. Equations and graphs show how much shorter the new guessing-free test can be and still retain the original measurement validity.

Additional equations and curves show that choice tests with zero measurement validities can have appreciable reliabilities due to differences in guessing strategy in the population.

All the analyses indicate that conversion to admissible probability measurement will yield quite significant improvements in measurement validity along with considerable reductions in test length.

DD FORM 1473

UNCLASSIFIED

Security Classification

test reliability
test validity
length of test
guessing
test wiseness
admissible probability measurement